Non-Reversible Parallel Tempering on Optimized Paths

by

Saifuddin Syed

B.Math., The University of Waterloo, 2014
M.Sc., The University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2022

The following individuals certify that they recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Non-Reversible Parallel Tempering on Optimized Paths

submitted by **Saifuddin Syed** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Statistics**.

Examining Committee:

Alexandre Bouchard-Côté, Associate Professor, Department of Statistics, University of British Columbia

Supervisor

Trevor Campbell, Assistant Professor, Department of Statistics, University of British Columbia

Supervisory Committee Member

Benjamin Bloem-Reddy, Assistant Professor, Department of Statistics, University of British Columbia

Supervisory Committee Member

Matías Salibián-Barrera, Professor, Department of Statistics, University of British Columbia

University Examiner

Daniel Coombs, Professor, Department of Mathematics, University of British Columbia

University Examiner

Pierre Jacob, Professor, Département des Systèmes d'Information, Sciences de la Décision et Statistiques, ESSEC Business School

External Examiner

Abstract

Parallel tempering (PT) methods are a popular class of Markov chain Monte Carlo schemes used to sample complex high-dimensional probability distributions. They rely on a collection of N interacting auxiliary chains targeting tempered versions of the target distribution to improve the exploration of the state-space. We provide here a new perspective on these highly parallel algorithms and their tuning by identifying and formalizing a sharp divide in the behaviour and performance of reversible versus non-reversible PT schemes.

We show theoretically and empirically that a class of non-reversible PT methods dominates its reversible counterparts. These results are exploited to identify the optimal annealing schedule for non-reversible PT and to develop an iterative scheme approximating this schedule. The proposed methodology is applicable to sample from a distribution π_1 with respect to a reference distribution π_0, and to compute normalizing constants. We provide a wide range of numerical examples supporting our theoretical and methodological contributions.

The performance of non-reversible PT depends on how quickly a sample from the reference distribution makes its way to the target, which in turn depends on the particular path of annealing distributions. Traditionally PT has used only simple paths constructed from convex combinations of the reference and target log-densities; we demonstrate that this path performs poorly in the setting where the reference and target are nearly mutually singular. To address this issue, we expand the framework of PT to general families of paths, formulate the choice of a path as an optimization problem that admits tractable gradient estimates, and propose a flexible new family of spline interpolation paths for use in practice. We show that PT induces a geometry on the space of probability distributions and characterize these optimal paths as length minimizing geodesics between π_0 and π_1. Theoretical and empirical results demonstrate that our proposed methodology breaks previously-established upper-performance limits for traditional linear paths.

Finally, we identify distinct scaling limits for the non-reversible and reversible schemes, the

former being a piecewise-deterministic Markov process and the latter a diffusion.

Lay Summary

Markov Chain Monte Carlo (MCMC) is a powerful tool to model and simulate probabilistic systems. In practice, MCMC methods can fail to efficiently explore the entirety of a given model space of parameters and become trapped in local regions of high probability, limiting their use when tackling complex problems. Parallel tempering (PT) improves the accuracy of MCMC by deploying multiple interacting copies of MCMC simultaneously and encouraging better exploration of the parameters. The current paradigm for PT presumes a notion called reversibility, where the performance deteriorates when too many additional copies are introduced. Reversibility makes PT notoriously sensitive to tune and renders it unsuitable for the challenging problems we are faced with today. We propose a new non-reversible PT paradigm that dominates its reversible counterpart and improves performance with increased parallelism. We develop an efficient, general-purpose methodology around non-reversibility that scales to modern computational resources and can tackle large-scale problems.

Preface

This dissertation is the original work of Saifuddin Syed, completed under the supervision of Alexandre Bouchard-Côté.

Chapters 2, 3, and 5 are based on a paper co-authored with Professor Alexandre Bouchard-Côté, Professor George Deligiannidis, and Professor Arnaud Doucet titled "Non-reversible parallel tempering: A scalable highly parallel MCMC scheme" published in the Journal of the Royal Statistical Society Series B in 2021. The ideas for Chapter 2, and 3 have been jointly developed and refined by Saifuddin Syed, Professor Alexandre Bouchard-Côté, and Professor Arnaud Doucet. Saifuddin Syed developed the theoretical framework and wrote the manuscript as lead author. Saifuddin Syed and Alexandre Bouchard-Côté jointly developed the methodological framework. Alexandre Bouchard-Côté conducted the experiments in Chapter 3, especially those in Section 3.4. Saifuddin Syed developed the ideas in Chapter 5 with significant contribution from George Deligiannidis in proving the scaling limit of the scaled index process. Edwin Perkins also offered very insightful discussions for ideas in Chapter 5.

Chapter 4 is based on a paper co-authored with Vittorio Romaniello, Professor Trevor Campbell, and Professor Alexandre Bouchard-Côté titled "Parallel tempering on optimized paths" published in the Proceedings of the 38th International Conference on Machine Learning (ICML) in 2021. Saifuddin Syed developed the idea for this paper, and the theory with significant contributions from all co-authors. As lead co-authors, Vittorio Romaniello and Saifuddin Syed developed the path tuning algorithm and construction of the exponential annealing family. Vittorio Romaniello did the experimental work and created the optimization tools for path tuning; this material is not included in this dissertation. Chapter 4 is built on top of the ICML paper and has been heavily re-written by Saifuddin Syed. Section 4.5 is solely developed by Saifuddin Syed for this thesis and is not yet published.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my supervisor Alexandre Bouchard-Côté. Alex gave me the freedom to wander with the support and guidance ensuring I would not get lost. Our conversations blurred the line between work and play and left me inspired to create.

I would also thank my mentors Trevor Campbell, Benjamin Bloem-Reddy, and Arnaud Doucet for guiding me throughout my career. I am a more capable researcher and mathematician because of them. Trevor challenged me to venture out of my comfort zone, provided me with new tools, and taught me how to sharpen them. Ben encouraged me to prioritize beauty in my work and adventure in my life. Arnaud gifted me his time and helped me build confidence in my ideas and future. I am excited to see where the next few years will take us.

I owe a debt of gratitude to my collaborators, George Deligiannidis and Vittorio Romaniello, who helped me unlock the potential of my work. I want to thank everyone in Alex and Trevor's reading group for allowing me to learn from them and exposing me to new ideas that lead to this dissertation. I also want to apologize to everyone in Alex and Trevor's reading group for being forced to endure my ramblings about the ideas that lead to this dissertation.

I would not have been able to do this without my army of a support system: Jonathan Agyeman, Jag Athwal, Maxime Bergeron, Creagh Briercliffe, Joshua Bon, Thom Bohdanowicz, Chanel Blouin, Sam Barney, Anthony Caterini, Natasha Camille, Elliot Cheung, Jonathan Alexander Coates, Cam Colleypriest, Michelle Chrabolowski, Katie Dixon, Thomas Hughes, Herby Desriveaux, Kat Dobson, Annie Dufficy, Catherine Ewasiuk, Luis Goddyn, Aaron Goodbaum, Dave Gu, Ali Hauschildt, Megan Ho, Thomas Hughes, Tom Hutchcroft, Isabela Jerônimo do Ó, Henry John, Christa Jeanne Johnson, Grant de Jong, Amit Kadan, Seth Kalyan, Santeri Kaupinmäki, Emma McCleod, Emma MacFarlane, Kenny Lobb, William Ou, Natasha Parent, David Perrin, Kevin Pierce, Charlie Payne, Beata Maksimowski, Vaden Masrani, Rudi Plesch, Emma Russo, PJ Salehi, Sohrab Salehi, Leslie Saunders, Bobak Shahriari, Ian Shaw, Jaclyn Shirley, Hannah Shumka, Liam Siemens,

Andrea Sollberger, Kate Sullivan, Madison Thulien, Paul Tiede, Sam Viavant, Ben Wallace, Grant Watson, Joe Watson, Alena Webber, Mary Wilson, Shell Wyngarde, Sean Xu, plus everyone in the Commercial Drive crew, Loose Joints, Lost in the Moss, Paper Crane, and Tuesday running group. An extra special shout-out to my therapist Lindsey Jepperson. This is not an exhaustive list. Everyone here provided me with a significant amount of academic and personal support during my Ph.D. You championed me and filled my life with conversation, laughter, colour, and adventure.

Finally, I want to thank my family. My sisters and brothers, Namirah Anis, Maryum Anis, Devin Maguire, and Varun Poduval, for having more faith in me than I ever did. To my parents, for being my anchors and ensuring that I knew I was never alone and unconditionally loved every day.

Ami and Abu,

Your immeasurable love and sacrifice gifted me a world full of beauty and knowledge.

Chapter 1

Introduction

If one ox could not do the job they did not try to grow a bigger ox, but used two oxen. When we need greater computer power, the answer is not to get a bigger computer, but... to build systems of computers and operate them in parallel.

— Admiral Grace Hopper

1.1 Markov Chain Monte Carlo

Monte Carlo methods are a set of tools used to numerically simulate random systems and used to solve a large class of challenging problems in statistics and science. Monte Carlo methods gamble the accuracy of the solution at the expense of a fixed computational budget. They were born during the Manhattan Project in Los Alamos, New Mexico, by physicists Stanislaw Ulam and John von Neumann to compute the probability of winning solitaire and study thermonuclear fission (Metropolis and Ulam, 1949).

Markov Chain Monte Carlo (MCMC) is a general-purpose class of Monte Carlo methods to simulate from general probability distributions. The first MCMC algorithm was also developed by physicists at Los Alamos to simulate statistical mechanical systems (Metropolis et al., 1953). It was later generalized and introduced to statisticians by Keith W. Hastings as the famous "Metropolis-Hastings" algorithm (Hastings, 1970). Due to limitations in theoretical knowledge and computing power at the time, it was not easy to realize the potential of MCMC algorithms and they were subsequently largely ignored by statisticians.

Simultaneously, statisticians were developing a theory of statistical modelling using Bayes theorem as an alternative to the frequentist philosophy prevalent at the time. The central object in Bayesian statistics is the *posterior* distribution π_1 defined over some state-space of statistical models $\mathcal{X}$. It is generally possible to evaluate the posterior density of $\pi_1(x)$ up to a normalizing

constant but one cannot simulate from the posterior directly. The computational challenges in simulating from the posterior were one of the limitations inhibiting the adoption of the Bayesian philosophy. As computing power became cheaper, Bayesian statisticians saw the potential MCMC as a tool for posterior inference (Gelfand and Smith, 1990). The discovery that MCMC could enable posterior inference led to a revolution in statistics. In the words of McGrayne (2011): "The combination of Bayes and MCMC has been called 'arguably the most powerful mechanism ever created for processing data and knowledge.' Almost instantaneously, MCMC changed statisticians' entire method of attacking problems. MCMC solved real problems, used computer algorithms instead of theorems, and led statisticians and scientists into a world where 'exact' meant 'simulated' and repetitive computer operations replaced mathematical equations."

Formally, the goal in Bayesian statistics is to make inferences using the posterior distribution π_1 by computing quantities in the form of an expectation,

$$\mathbb{E}[f] := \int_{\mathcal{X}} f(x)\pi_1(x)\mathrm{d}x, \tag{1.1}$$

where π_1 is the *posterior distribution* over some space of statistical models $\mathcal{X}$, and $f : \mathcal{X} \to \mathbb{R}$ is some quantity of interest about the model. The main idea behind MCMC is to construct a π_1-stationary Markov chain X_t that efficiently explores the state-space $\mathcal{X}$ while preserving the statistics of the target π_1. We then take an average over the trajectory of the path in $\mathcal{X}$ to approximate (1.1) at the price of a statistical error, by invoking the law of large numbers for Markov chains,

$$\mathbb{E}[f] \overset{a.s.}{=} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t). \tag{1.2}$$

Unfortunately, the rate of convergence for (1.2) can often be poor for challenging problems where π_1 has multiple well-separated modes with highly varying curvature, or when $\mathcal{X}$ is high dimensional with topological constraints. In these situations, the chain gets trapped exploring a local region of high probability, failing to mix within the computational budget and resulting in a poor reconstruction of the target π_1 (see Figure 1.1 (top)). In practice, a practitioner does not know the structure of the posterior for challenging problems, making it difficult to assess when the Markov chain has converged or has gotten stuck due to poor exploration. Designing efficient and

robust MCMC methodology where (1.2) reliably converges within the computational constraints is of fundamental importance to the adoption of Bayesian statistics.

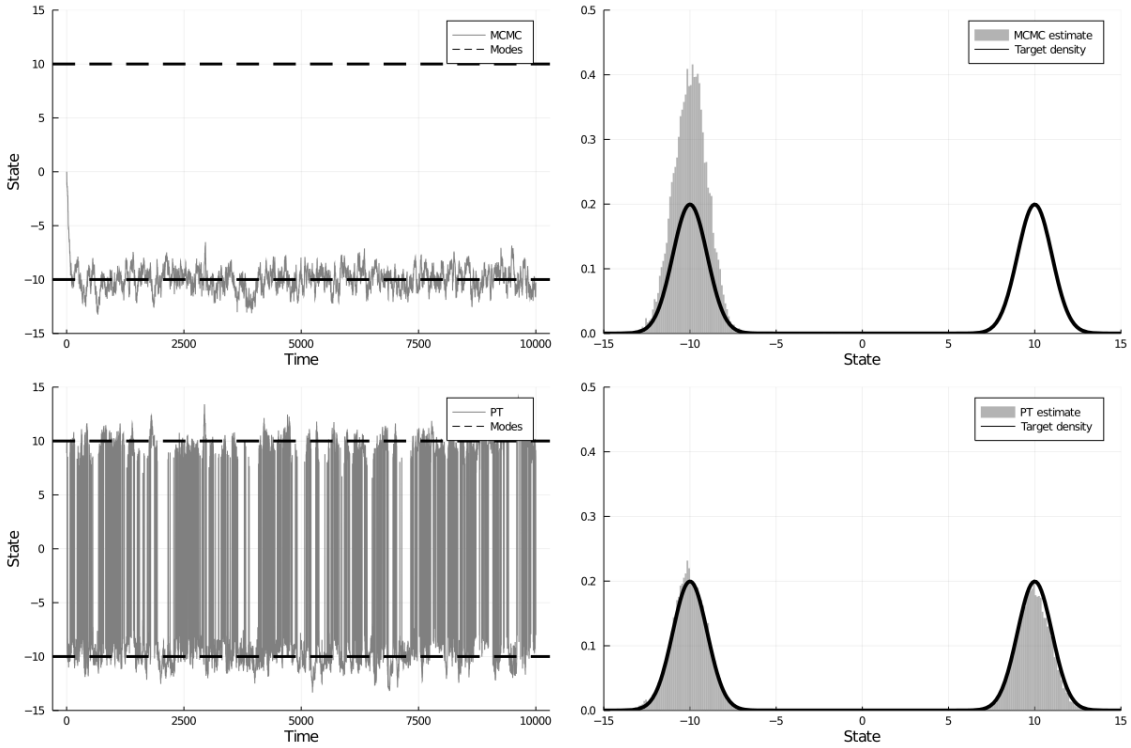

Figure 1.1: Trace plots (left) and density estimates (right) of MCMC versus PT when π_1 is a mixture of $N(-10, 1)$ and $N(10, 1)$ with well-separated modes at $x = -10$ and 10. (Top) The trajectory of a single MCMC chain built using random walk Metropolis algorithm gets trapped exploring one mode and fails to converge. (Bottom) PT trajectories used the same local exploration moves with $N = 10$ chains and a standard Gaussian reference. Not only did the PT chain discover both modes, but it also easily traversed between them.

One natural direction to handle these complex problems is using multiple processors or distributing the computation. Since the demise of Moore's law in the past decade, MCMC methods must be designed to take advantage of modern computational architectures such as GPUs and distributed computing. There have been attempts made to improve the scalability of MCMC using parallel computing in the "big data" regime by exploiting parallel processors within each iteration (e.g. see Jacob et al. (2011); Brockwell (2006); Lee et al. (2010); Calderhead (2014); Scott et al. (2016); Wang et al. (2015); Wu and Robert (2017); Bardenet et al. (2017); Jacob et al. (2020)). Most of these methods are motivated by applications, where memory is constrained, or the size of the data and model make the computation cost of MCMC is prohibitively expensive. Still, they do not address the challenges faced with problems involving "big models," where the bottleneck is not memory

but rather the complexity of the target π_1 and the state-space $\mathcal{X}$ which we have to explore. We are interested here in this latter category which includes challenging problems such as cosmological modelling, phylogenetic inference, spin glasses, space-time random effect models, protein folding, or multiple sequence alignment.

1.2 Parallel tempering

A popular approach for multi-core and distributed exploration of complex distributions is Parallel Tempering (PT) (also known as the Replica Exchange Method (REM), and Metropolis-Coupled Markov Chain Monte Carlo (MC^3)) which was introduced independently in statistics (Geyer, 1991) and physics (Hukushima and Nemoto, 1996); see also Swendsen and Wang (1986) for an earlier related proposal. To sample from the target distribution π_1, PT introduces a sequence of auxiliary *tempered* or *annealed* probability distributions with densities

$$\pi_{\beta_n}(x) \propto \pi_0(x)^{1-\beta_n}\pi_1(x)^{\beta_n}, \quad n = 0, 1, ..., N,$$

where π_0 is an easy-to-sample reference distribution, and the sequence

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1,$$

defines the *annealing schedule*. This bridge of auxiliary distributions $\pi_{\beta_0}, \pi_{\beta_1}, \ldots, \pi_{\beta_N}$ is used to progressively transform samples from the *reference distribution* ($\beta = 0$) into samples from the *target distribution* ($\beta = 1$), for which only poorly mixing MCMC kernels may be available (see Figure 1.2). For example, in the Bayesian setting where the target distribution is the posterior, we can choose the reference distribution as the *prior*, from which we can often obtain independent samples.

More precisely, PT algorithms construct a Markov chain $\mathbf{X}_t = (X_t^0, \ldots, X_t^N)$ in which the states are $(N+1)$-tuples, $\boldsymbol{x} = (x^0, x^1, x^2, \ldots, x^N) \in \mathcal{X}^{N+1}$, and whose stationary distribution is given by $\boldsymbol{\pi}(\boldsymbol{x}) = \prod_{n=0}^{N} \pi_{\beta_n}(x^n)$ (Geyer, 1991). At each iteration, PT proceeds by applying in parallel $N+1$ MCMC kernels targeting π_{β_n} for $n = 0, ..., N$. We call these model-specific kernels the *local exploration kernels* (Figure 1.3 (middle)). The chains closer to the reference chain (i.e. those with annealing parameter β_n close to zero) can typically traverse regions of low probability mass under π_{β_n}

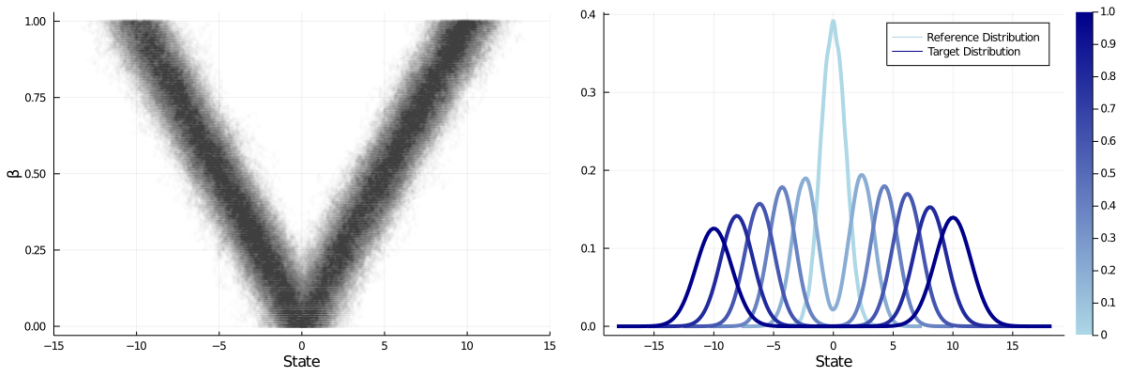

Figure 1.2: Example of an annealing path $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^\beta$ of distributions when the target π_1 is a mixture of Gaussian $N(-10, 1)$ and $N(10, 1)$, and the reference π_0 is the standard Gaussian $N(0, 1)$. (Left) The density of π_β continuously changes from π_0 to π_1 as β increases from 0 to 1. The annealing distributions uses the reference distribution to create a tunnel between the two modes of the target. (Right) The annealing distributions discretize this transition between the reference and target.

while the chain $\beta_N = 1$ ensures that asymptotically we obtain samples from the target distribution. 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. A *communication move* corresponds to a collection of swap moves (Figure 1.3 (right)). A maximal collection of non-interfering swaps are the *Even* and *Odd swaps*, which propose to exchange states at chains with index n and $n+1$ with an even and odd index n respectively. We will refer to a local exploration move followed by a communication move as a *scan* which can be visualized in Figure 1.3.

Based on the assumption that the reference distribution π_0 can be sampled efficiently, PT uses the swap-based interactions between neighbouring chains to propagate the exploration done in π_0 into improved exploration in the chain of interest π_1. PT guarantees that the marginal distribution of the N^{th} chain converges to π_1. By using the law of large numbers for Markov chains on the N-th component of $\mathbf{X}_t$, we can recover almost surely,

$$\mathbb{E}[f] \overset{a.s.}{=} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t^N).$$

In practice, the rate of convergence is often much faster compared to running a single chain (Woodard

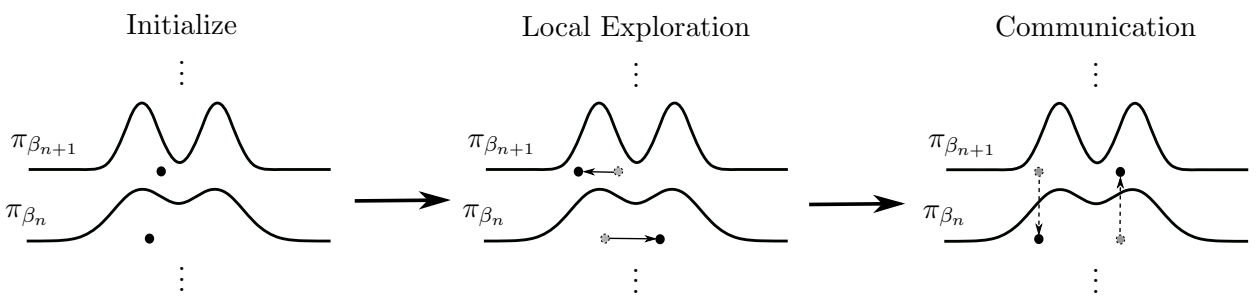

Figure 1.3: Visualization of a scan consisting of a local exploration move (middle) and a communication move (right). (Left) The states are initialized from π_{β_n} and $\pi_{\beta_{n+1}}$, respectively. (Middle) Each state independently explores the state-space during the local exploration move according to its corresponding annealing distribution. The states traverse modes relatively easily for π_{β_n} and struggle for $\pi_{\beta_{n+1}}$. (Right) Nearest neighbour chains communicate through a proposed swap of their states. In this case, the final state for chain $n+1$ after a successful swap is located in a different mode of $\pi_{\beta_{n+1}}$ than initialized.

et al., 2009). In particular, Woodard et al. (2009) showed that when a single chain targeting π_1 mixes poorly, the global PT chain will mix rapidly if the follow reasonable conditions are met: (1) the exploration kernel mixes well at the reference, (2) the exploration kernel mixes well within a mode for each β_n, and (3) $\pi_{\beta_{n-1}}$ and π_{β_n} have sufficient overlap in their modes for $n = 1, \ldots, N$.

Therefore, parallel tempering can be seen as a "meta-algorithm" that uses parallel computing to transform a locally well-mixing target chain into a globally well-mixing chain. In particular, it does not presume any structural assumption on the state-space or the target π_1, making it a general-purpose tool for Bayesian inference. PT remains to this day a widely used MCMC method to sample from complex multimodal target distributions arising in physics, chemistry, biology, statistics, and machine learning; see, for example, Issaoun et al. (2021); Ballnus et al. (2017); Chandra et al. (2019); Cho et al. (2010); Desjardins et al. (2014); Diaz et al. (2020); Dorri et al. (2020); Kamberaj (2020); Müller and Bouckaert (2020).

1.2.1 Non-reversible parallel tempering

A key notion used to analyze the behaviour of PT is the *index process*. To provide intuition on this process, it is helpful to discuss briefly how PT is distributed over several machines. An important point is that instead of having pairs of machines exchanging high-dimensional states when a swap is accepted (which could be detrimental due to network latency), the machines should just exchange

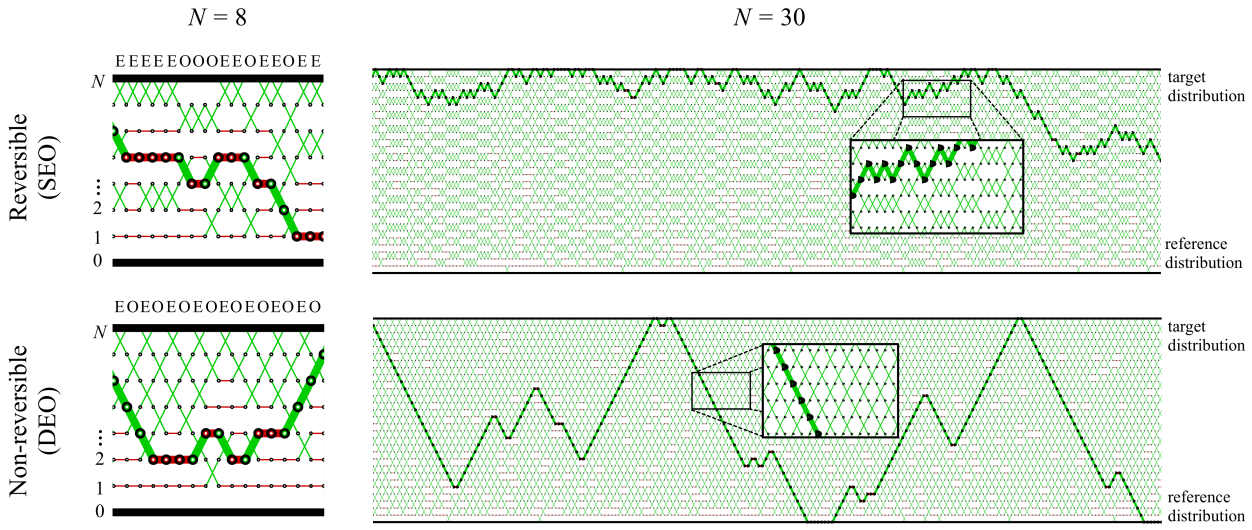

Figure 1.4: Index process for reversible (top) and non-reversible (bottom) PT for $N = 8$ (left) and $N = 30$ auxiliary chains (right) using equally spaced annealing parameters on a Bayesian change-point detection model (Davidson-Pilon, 2015) where π_0 is the prior, π_1 the posterior. The even and odd swaps moves are labeled "E", and "O" respectively. The sequence of swap moves forms $N + 1$ index process trajectories (paths formed by the red and green edges). We show one such path in bold. The reversible and non-reversible PT clearly exhibit different scaling behaviour which we formalize in Section 5.2.

the annealing parameters. Suppose now we initialize machine m with annealing parameter β_m. Then after t scans, the annealing parameters are permuted among the $N + 1$ machines according to the permutation $\mathbf{I}_t = (I_t^0, \ldots, I_t^N)$ of $\{0, \ldots, N\}$, so that machine m has annealing parameter $\beta_{I_t^m}$ at iteration t of PT. Each *index process* $(I_t^m)_{t=1}^{\infty}$, formally introduced in Section 2.3, is initialized using $I_0^m = m$ and tracks how the state of the corresponding chain evolves over the annealing schedule thanks to the swap moves; see Figure 1.4. The index process I_t^m thus monitors the information transfer between the reference and target on machine m and as such determines partly the effectiveness of PT.

There have been many proposals made to improve this information transfer by adjusting the annealing schedule; see, e.g., Kone and Kofke (2005); Atchadé et al. (2011); Miasojedow et al. (2013). These proposals are useful but do not address a crucial limitation of standard PT algorithms. In a distributed context, one can select randomly at each iteration whether to apply Even or Odd swap moves from Section 1.2 in parallel . The resulting stochastic even-odd swap (SEO) scheme (Lingenheil et al., 2009), henceforth referred to as *reversible PT* as it admits a reversible scaling limit (see Section 5.2), yields index processes exhibiting a diffusive behaviour; see top row of Figure

1.4. 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 $\beta = 1$ (Diaconis et al., 2000). 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 the reference and target to communicate. 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 Section 3.1.4 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 squared jump distance travelled per swap in the space of annealing parameters $[0, 1]$ (Rathore et al., 2005; Kone and Kofke, 2005; Lingenheil et al., 2009; Atchadé et al., 2011). Adding more chains past this threshold actually deteriorates the performance of reversible PT and there have even been attempts to adaptively reduce the number of additional chains (Łacki and Miasojedow, 2016). This is a lost opportunity, since PT is otherwise particularly suitable to implementation on multi-core or distributed architectures.

An alternative to the SEO scheme is the deterministic even-odd swap (DEO) scheme introduced in Okabe et al. (2001); Lingenheil et al. (2009) where one deterministically alternates Even and Odd swap moves. We refer to DEO as *non-reversible PT* as it admits a non-reversible scaling limit (see Section 5.2). In particular, the resulting index processes do not appear to exhibit a diffusive, i.e. random walk type, behaviour, illustrated by the bottom row of Figure 1.4. This non-diffusive behaviour of non-reversible PT explains its excellent empirical performance when compared to alternative reversible PT schemes observed in practice (Lingenheil et al., 2009) for any given schedule. However non-reversible PT, like reversible PT, is sensitive to the choice of the schedule. All the aforementioned tuning strategies developed for PT implicitly assume a reversible framework and do not apply to non-reversible PT (as empirically verified in Lingenheil et al. (2009)).

1.3 Outline of thesis

The purpose of this thesis is to identify some of the theoretical properties of non-reversible PT so as to establish optimal tuning guidelines for this algorithm and propose novel methodology for implementation.

1.3.1 Chapter 2

In chapter 2, we formally introduce the PT algorithm with both reversible and non-reversible communication and provide the relevant background information for the remaining chapters. In particular, we introduce a dual perspective of parallel tempering in a distributed computing framework, which motivates the construction of the index process. The index process tracks the communication between the reference and target on each machine and will be the focus of both the theoretical and methodological development of PT. Finally, we motivate the round trip rate as the natural the notion of optimality for PT, which quantifies how often information from the reference distribution percolates to the target through the index process.

1.3.2 Chapter 3

Our first contribution is a non-asymptotic result showing that the non-reversible DEO scheme is guaranteed to outperform its reversible SEO counterpart. The non-asymptotic analysis of the round trip rate is based on a simplifying assumption called Efficient Local Exploration (ELE) and shows that the round trip rate for non-reverible PT dominates its reversible counterpart in all scenarios (see Corrollary 2). We do not expect ELE to hold exactly in real scenarios, however we show empirically that there are practical methods to approximate it. Even when ELE is violated the key predictions made by the theory closely match empirical behaviour. In this sense ELE can be thought of as a useful model for understanding PT algorithms.

In Section 3.2 we introduce the local and global communication barriers $\lambda(\beta), \Lambda$ which encode the local and global efficiency of PT respectively. We then show that for non-reversible PT the round trip rate converges to $(2 + 2\Lambda)^{-1}$, in contrast to the reversible counterpart for which it decays to zero. To provide some intuition on how the small algorithmic difference between SEO and DEO can have such a profound impact, consider the scenario where PT would use the same distributions at the two end-points, $\pi_0 = \pi_1$, as well as for all intermediate distributions so that $\Lambda = 0$. Clearly this is not a realistic scenario, but in this simple context the index processes of DEO and SEO are easy to describe and contrast. In both DEO and SEO, $\pi_0 = \pi_1$ implies that all proposed swaps will be accepted. For DEO, this makes the index process fully deterministic, performing direct trips from index 0 to N and back. Such a process could be compared to a "conveyor belt" with the

property that no matter what is the value of N, one novel trip from chain 0 reaches chain N every two iterations, i.e. the round trip rate is $(2 + 2\Lambda)^{-1} = 1/2$ as $\Lambda = 0$. For SEO, even when $\pi_0 = \pi_1$ the index process is still random and can be readily seen to be a simple discrete random walk. This implies that as N increases, the round trip rate decreases to zero.

In practice, achieving high round trip rates requires careful tuning of the annealing schedule. In Section 3.3 we combine the analysis from Section 3.1 and Section 3.2 to develop a novel methodology to optimize the annealing parameters and optimally allocate of computational resource. The optimal tuning guidelines we provide are different from existing (reversible) PT guidelines and the novel methodology is highly parallel as its performance does not collapse when a very large number of chains is used. However using a large number of chains does have a diminishing return, therefore we propose a mechanism to determine the optimal trade-off between the number of chains and the number of independent PT algorithms one should use.

Finally in Section 3.4, we present a variety of experiments validating our theoretical analysis and novel methodology. These examples include nine Bayesian models ranging from simple foundational models such as generalized linear models and Bayesian mixture, to complex ones such as cancer copy-number calling, ODE parameter estimations, spike-and-slab classification and two types of phylogenetic models. This is complemented by three popular models in statistical mechanics and four artificial models. These experiments also include eight real datasets spanning diverse data-types and size including modern types of measurements such as whole-genome single-cell sequencing data (494 individual cells from two types of cancer, triple negative breast cancer and high-grade serous ovarian) and mRNA transfection time series (Dorri et al., 2020; Leonhardt et al., 2014), as well as more conventional ones such as primate mitochondrial DNA data and various feature selection/classification datasets.

The method is implemented in an open source Bayesian modelling language available at `https://github.com/UBC-Stat-ML/blangSDK`. Our software implementation allows the user to specify the model in a BUGS-like language (Lunn et al., 2000). From this model declaration, a suitable sequence of annealed distributions is instantiated and a schedule optimized using our iterative method.

1.3.3 Chapter 4

Chapter 3 showed that the optimal performance of PT is a governed by the communication barrier Λ, which depends on the path between the reference and target. In particular, increasing the number of chains or tuning the annealing parameters cannot improve the optimal round trip rate. So far in our analysis, we have presumed $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^{\beta}$ is the *linear annealing path* constructed from convex combinations of the reference and target log-densities. The linear annealing path has deep roots in physics and information theory via Gibbs distributions and is traditionally used in computational statistics for its simplicity (Neal, 2001; Del Moral et al., 2006); however, it can be sub-optimal for PT (Tawn et al., 2020).

Chapter 4 begins by demonstrating that the linear path performs poorly in the setting where the reference and target are nearly mutually singular. To address this issue, in Section 4.2 we expand the framework of PT to general non-linear annealing paths π_β and show that the non-asymptotic and asymptotic analysis from Chapter 3 naturally extends to PT on general annealing paths. In particular, the asymptotic round trip rate still converges to $(2+2\Lambda)^{-1}$, where now the communication barrier Λ is a function of the path (see Section 4.3). In Section 4.4 we formulate the choice of the annealing path as an optimization problem that admits tractable gradient estimates.

In Section 4.5, we explore the geometric properties of annealing in the context of PT. When annealing paths take values in a parametric family of distributions, we show that PT induces a natural geometry where the local and global communication barriers are the speed and length of the path, respectively. This geometric view provides intuition for tuning PT: the optimal schedule as a constant speed reparameterization, and the optimal path as the geodesic minimizing the length between π_0 and π_1. When π_0 and π_1 are nearly-mutually singular, we show that appropriately tuning the path can lead to exponential reductions in Λ compared to the traditional path.

We propose a natural parametric family constructed from π_0 and π_1 that is compatible with parallel tempering. It admits a flexible new family of spline interpolation paths for use in practice. We provide theoretical and empirical results to demonstrate that our proposed methodology breaks the previously-established upper-performance limits for the linear path.

1.3.4 Chapter 5

In Chapter 5.2 we identify the scaling limit of the index processes for both reversible and non-reversible PT as the number of parallel chains goes to infinity. We characterize the scaled index processes through their infinitesimal generators, and we show rigorously that the scaling limit is a piecewise-deterministic Markov process for non-reversible PT. In contrast, it is a diffusion for reversible PT as suggested by the dynamics of the bold paths in Figure 1.4. This divergence in scaling laws helps explain how such an innocuous change between the reversible and non-reversible PT algorithms can substantially impact performance.

Chapter 2

Parallel Tempering

We stand at the threshold of a many core world.
The hardware community is ready to cross this threshold.
The parallel software community is not.
— Tim Mattson

2.1 Annealing

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Denote $\mathcal{P}(\mathcal{X})$ be the set of probability densities with full support on a state space $\mathcal{X}$ with respect to a common dominating measure $\mathrm{d}x$. For each $p \in \mathcal{P}(\mathcal{X})$, we will assume $p(x)$ the density with respect to $\mathrm{d}x$ can be evaluated up to a normalizing constant. Specifically, we presume the existence of $W : \mathcal{X} \to \mathbb{R}$ that we can efficiently evaluated $W(x)$ for each x and,

$$p(x) = \frac{1}{Z} \exp(W(x)),$$

where $Z = \int_{\mathcal{X}} \exp(W(x))\mathrm{d}x < \infty$. Given $f : \mathcal{X} \to \mathbb{R}$ is integrable with respect to $p \in \mathcal{P}(\mathcal{X})$, we will define the expectation of $\mathbb{E}_p[f]$ to be the expectation of f with respect to p,

$$\mathbb{E}_p[f] = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x.$$

With the aid of a *reference* distribution $\pi_0 \in \mathcal{P}(\mathcal{X})$ which we can efficiently sample from and evaluate $\pi_0(x)$, our goal will be to compute $\mathbb{E}_{\pi_1}[f(X)]$ for some *target distribution* $\pi_1 \in \mathcal{P}(\mathcal{X})$ with log likelihood $W_1(x)$.

2.1.1 The linear annealing path

We define the *linear annealing path* between $\pi_0 \propto \exp(W_0)$ and $\pi_1 \propto \exp(W_1)$ as $\pi : [0,1] \to \mathcal{P}(\mathcal{X})$ denoted $\beta \mapsto \pi_\beta \propto \exp(W_\beta)$, with log-densities $W_\beta = (1-\beta)W_0 + \beta W_1$ linearly interpolating between W_0 and W_1. The *annealing distribution* π_β for the linear path satisfy,

$$\pi_\beta(x) = \frac{1}{Z(\beta)} \exp(W_0 + \beta V(x)), \quad x \in \mathcal{X}$$

where $Z(\beta) = \int_{\mathcal{X}} \exp\left(W_\beta(x)\right) \mathrm{d}x$ is the normalizing constant and $V(x) = W_1(x) - W_0(x)$ is the log-likelihood ratio between π_1 and π_0 modulo a constant factor. We will assume we can efficiently evaluate $V(x)$ for $x \in \mathcal{X}$ but not $Z(\beta)$ for $\beta > 0$. See Figure 1.2 for an example of a linear path.

To simplify notation, given $f : \mathcal{X} \to \mathbb{R}$ integrable with respect to π_β, define $\mathbb{E}_\beta[f] := \mathbb{E}_{\pi_\beta}[f]$ to be the expectation of f with respect to π_β,

$$\mathbb{E}_\beta[f] = \int_{\mathcal{X}} f(x)\pi_\beta(x)\mathrm{d}x. \tag{2.1}$$

In particular, we are interested in approximating (2.1) at $\beta = 1$.

In the statistical physics literature the π_β commonly referred to as the *Gibbs distribution* at *inverse-temperature* β and $-V(x)$ and $Z(\beta)$ are the *potential energy* and *partition function* respectively.

2.1.2 Annealing schedule

We denote the *annealing schedule* as a partition $\mathcal{B}_N = (\beta_0, \ldots, \beta_N)$ of $[0,1]$ where

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1,$$

with mesh size $\|\mathcal{B}_N\| = \sup_n \ |\beta_n - \beta_{n-1}|$.

Suppose $\gamma : [0,1] \to [0,1]$ is a strictly increasing differentiable function satisfying $\gamma(0) = 0$ and $\gamma(1) = 1$. We say that γ is a *schedule generator* for $\mathcal{B}_N = (\beta_0, \ldots, \beta_N)$ if $\beta_n = \gamma(n/N)$. In particular

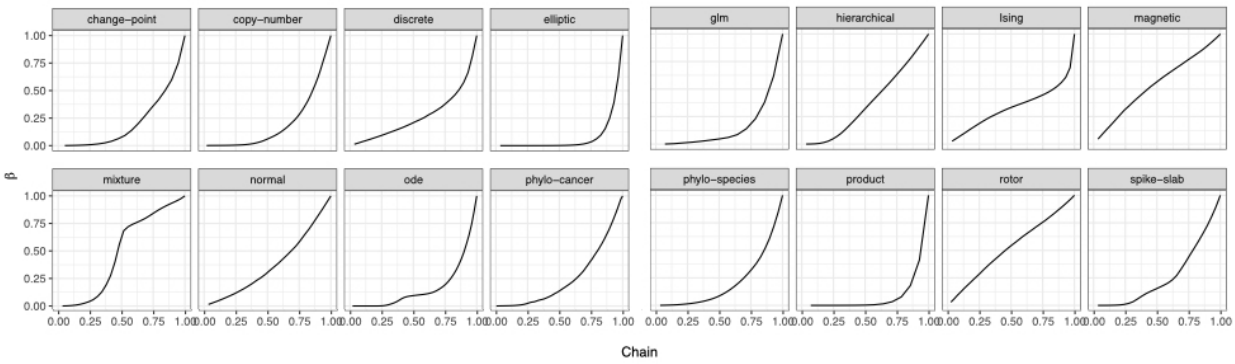

Figure 2.1: Estimates of the schedule generator γ for 16 models (see Section 3.4.1 for details). The abscissa denotes the normalized chain indices n/N, the ordinates, parameters β. The function γ is such that the schedule $\beta_n = \gamma(n/N)$ approximates equi-acceptance.

this implies,

$$\|\mathcal{B}_N\| \leq \frac{\|\dot{\gamma}\|_\infty}{N}, \tag{2.2}$$

where $\dot{\gamma}(w) = \frac{\mathrm{d}\gamma(w)}{\mathrm{d}w}$ and $\|\dot{\gamma}\|_\infty = \sup_w |\dot{\gamma}(w)|$ is the infinity norm.

Without loss of generality, we can always assume that $\mathcal{B}_N$ is generated by γ for all N. This is not as strict as it seems, since for a fixed N such a γ always exists. Moreover, most annealing schedules commonly used fall within this framework: for example, the uniform schedule $\mathcal{B}_N = (0, \frac{1}{N}, \ldots, 1)$ is generated by $\gamma(w) = w$. If $\pi_0(x) \propto \pi(x)^{\beta_*}$ for some $\beta_* \in (0, 1)$, then $\gamma(w) = \frac{\beta_*^{1-w} - \beta_*}{1-\beta_*}$ generates the *geometric schedule*.

In Section 3.3 in the next chapters we will determine what constitutes an optimal schedule $\mathcal{B}_N$ and how to efficiently learn its generator γ. We show in Figure 2.1 examples of schedule generators γ corresponding to estimated optimal schedule from various models. Figure 2.1 shows that in real world inference scenarios, the optimal schedule often qualitatively differs from the simple geometric schedule commonly used.

2.2 Parallel Tempering algorithm

We define the PT state $\boldsymbol{x} \in \mathcal{X}^{N+1}$ where $\boldsymbol{x} = (x^0, \ldots, x^N) \in \mathcal{X}^{N+1}$. The PT algorithm involves constructing a Markov chain $\mathbf{X}_t = (X_t^0, \ldots, X_t^N)$ over $\mathcal{X}^{N+1}$ invariant with respect to

$$\boldsymbol{\pi}(\boldsymbol{x}) = \prod_{n=0}^{N} \pi_{\beta_n}(x^n).$$

In particular, X_t^n tracks the *states* associated with annealing parameter β_n at scan t. In the literature it is common to refer to the sequence of states associated with β_n as the n-th *chain*.

2.2.1 Local exploration kernels

The local exploration kernels are defined in the same way for Stochastic Even Odd (SEO) and Deterministic Even Odd (DEO) communication schemes from Section 1.2.1. They are also model specific, so we assume we are given one π_{β_n}-invariant kernel K_{β_n} for each annealing parameter $\beta_n \in \mathcal{B}$. These can be based on t_{expl} steps of Hamiltonian Monte Carlo, Metropolis–Hastings, Gibbs sampling, slice sampling, etc. We construct the overall local exploration kernel by applying the annealing parameter specific kernels to each component independently from each other:

$$\mathbf{K}^{\text{expl}}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') = \prod_{n=0}^{N} K_{\beta_n}(x^n, \mathrm{d}x'^n).$$

See Figure 1.3 (middle) for a visualization of the local exploration kernel $\mathbf{K}^{\text{expl}}$.

In our computational model, we implicitly assume that the local exploration kernel at $\beta = 0$ is special in that it can provide independent exact samples from π_0. Mathematically, $K_0(x, A_0) = \pi_0(A_0)$. This assumption is satisfied in most Bayesian models equipped with proper prior distributions, but also in other situations such as Markov random fields (see Section 3.4.1).

2.2.2 Communication kernels.

Swaps

Before defining the communication scheme, we first construct its fundamental building block, a *swap*. A swap is a Metropolis–Hastings move with a deterministic proposal $\boldsymbol{x}^{(n,n+1)}$ for some

$n = 0, \ldots, N-1$ where

$$\boldsymbol{x}^{(n,n+1)} = (x^0, \ldots, x^{n-1}, x^{n+1}, x^n, x^{n+2}, \ldots, x^N),$$

consists of swapping n and $n+1$-th components of $\boldsymbol{x}$. The Metropolis-Hastings kernel $\mathbf{K}^{(n,n+1)}$ corresponding to this update is given by

$$\mathbf{K}^{(n,n+1)}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') = (1 - \alpha^{(n,n+1)}(\boldsymbol{x}))\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{x}') + \alpha^{(n,n+1)}(\boldsymbol{x})\delta_{\boldsymbol{x}^{(n,n+1)}}(\mathrm{d}\boldsymbol{x}').$$

The function $\alpha^{(n,n+1)}(\boldsymbol{x})$ is the corresponding acceptance probability equal to

$$\begin{aligned}\alpha^{(n,n+1)}(\boldsymbol{x}) &= 1 \wedge \frac{\boldsymbol{\pi}(\boldsymbol{x}^{(n,n+1)})}{\boldsymbol{\pi}(\boldsymbol{x})} \\ &= 1 \wedge \exp(\Delta W_{n+1}(x^n) - \Delta W_{n+1}(x^{n+1})),\end{aligned} \tag{2.3}$$

where $\Delta W_n = W_{\beta_n}(x) - W_{\beta_{n-1}}(x)$ is the change in log-density between π_{β_n} and $\pi_{\beta_{n-1}}$. For the linear path, ΔW_n simplifies to

$$\Delta W_n = (\beta_n - \beta_{n-1})V(x). \tag{2.4}$$

See Figure 1.3 (right) for a visualization of the swap kernel $\mathbf{K}^{(n,n+1)}$.

Odd/Even swaps

The maximal collection of adjacent swap kernels that can be proposed in parallel without interference are

$$\mathbf{K}^{\text{even}} := \prod_{n \text{ even}} \mathbf{K}^{(n,n+1)}, \qquad \mathbf{K}^{\text{odd}} := \prod_{n \text{ odd}} \mathbf{K}^{(n,n+1)},$$

which we call the *even* and *odd kernels* respectively.

For SEO, the kernel $\mathbf{K}_t^{\text{comm}} = \mathbf{K}^{\text{SEO}}$ is given by a mixture of the even and odd kernels in equal proportion while for DEO the kernel $\mathbf{K}_t^{\text{comm}} = \mathbf{K}_t^{\text{DEO}}$ is given by a deterministic alternation between

even and odd kernels, that is

$$\mathbf{K}^{\mathrm{SEO}} := \frac{1}{2}\mathbf{K}^{\mathrm{even}} + \frac{1}{2}\mathbf{K}^{\mathrm{odd}},$$

$$\mathbf{K}_t^{\mathrm{DEO}} := \begin{cases} \mathbf{K}^{\mathrm{even}} & \text{if } t \text{ is even,} \\ \mathbf{K}^{\mathrm{odd}} & \text{if } t \text{ is odd.} \end{cases}$$

2.2.3 PT kernel

For both SEO and DEO, the overall $\boldsymbol{\pi}$-invariant Markov kernel $\mathbf{K}_t^{\mathrm{PT}}$ describing the algorithm is obtained by the composition of a $\boldsymbol{\pi}$-invariant local exploration kernel $\mathbf{K}^{\mathrm{expl}}$ and communication kernel $\mathbf{K}_t^{\mathrm{comm}}$,

$$\mathbf{K}_t^{\mathrm{PT}}(\boldsymbol{x}, A) = \mathbf{K}_t^{\mathrm{comm}}\mathbf{K}^{\mathrm{expl}}(\boldsymbol{x}, A) := \int \mathbf{K}_t^{\mathrm{comm}}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')\mathbf{K}^{\mathrm{expl}}(\boldsymbol{x}', A).$$

The difference between SEO and DEO is in the communication phase, namely $\mathbf{K}_t^{\mathrm{comm}} = \mathbf{K}^{\mathrm{SEO}}$ in the former case and $\mathbf{K}_t^{\mathrm{comm}} = \mathbf{K}_t^{\mathrm{DEO}}$ in the latter. Markov kernels corresponding to the reversible (SEO) and non-reversible (DEO) PT algorithms are described informally in the introduction and illustrated in Figure 1.4. We define a *scan* to be one application of the PT kernel corresponding to one local exploration and communication kernel (see Figure 1.3).

We provide pseudo-code for the DEO scheme in Algorithm 1. The pseudo-code also estimates the average rejection probabilities $r^{(n,n+1)}$ of swap moves between chains $n-1$ and n which are used to optimize the annealing schedule in Section 3.3. When the schedule is fixed, lines 1, 14, 22 can be omitted, and one should use "for $n \in P$" on line 12. For simplicity, the swap in line 17 is shown for a *parallel computing* context, where several cores have a shared memory, and hence line 17 is simply an exchange of pointers in an array, which is efficient thanks to memory sharing.

2.3 Distributed PT

For a *distributed computing* implementation, where several machines do not share memory and instead need to communicate over the network, it becomes advantageous to swap annealing parameters instead of states. This motivates an alternative but equivalent view of PT by describing the dynamics

Algorithm 1 DEO

Input: Initial state $\boldsymbol{x}_0$, annealing path π, annealing schedule $\mathcal{B}_N$, number of scans t_{scan}

$r^{(n,n+1)} \leftarrow 0$ for all $n \in \{0, \dots, N-1\}$
▷ Initialize chain
$\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
for t **in** 1, 2, …, t_{scan} **do**
▷ Non-reversibility inducing alternation
  if t is even **then**
    $P \leftarrow \{n : 0 \leq n < N, n \text{ is even}\}$
  else
    $P \leftarrow \{n : 0 \leq n < N, n \text{ is odd}\}$
  end if
  ▷ Local exploration phase (parallelizable)
  for n **in** $0, \dots, N$ **do**
    $x^n \sim K_{\beta_n}(x^n_{t-1}, \cdot)$
  end for
  ▷ Communication phase (parallelizable)
  for n **in** $1, \dots, N$ **do**
  ▷ Compute acceptance probability using Equation (2.3).
    $\alpha \leftarrow \alpha^{(n,n+1)}$
    $r^{(n,n+1)} \leftarrow r^{(n,n+1)} + (1-\alpha)$
    $A \sim \text{Bern}(\alpha)$
    ▷ Chains swap states
    if $n \in P$ **and** $A = 1$ **then**
      $(x^{n-1}, x^n) \leftarrow (x^n, x^{n-1})$
    end if
  end for
  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}$
end for
▷ Compute mean rejection rate using Equation (3.18)
$r^{(n,n+1)} \leftarrow r^{(n,n+1)}/t_{\text{scan}}$ for all $n \in \{0, \dots, N-1\}$
return $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{t_{\text{scan}}}), (r^{(0,1)}, \dots, r^{(N-1,N)})$

taking place on each machine. Define the *replica process* for machine m as

$$(Y_t^m, I_t^m, \epsilon_t^m) \in \mathcal{X} \times \{0, \dots, N\} \times \{-1, 1\},$$

where at time t machine m stores state Y_t^m, annealing parameter $\beta_{I_t^m}$ and proposes a swap with the machine storing annealing parameter $\beta_{I_t^m + \epsilon_t^m}$. In particular $\mathbf{Y}_t = (Y_t^0, \dots, Y_t^N)$ is equivalent to $\mathbf{X}_t$ shuffled according to the *indices* $\mathbf{I}_t = (I_t^0, \dots, I_t^N)$ which permute $\{0, \dots, N\}$. The proposed evolution of the indices are governed by the *lifting parameters* $\boldsymbol{\epsilon}_t = (\epsilon^0, \dots, \epsilon_t^N)$. We can interpret I_t^m and ϵ_t^m as the "position" and "momentum" of the trajectory of the annealing parameters on

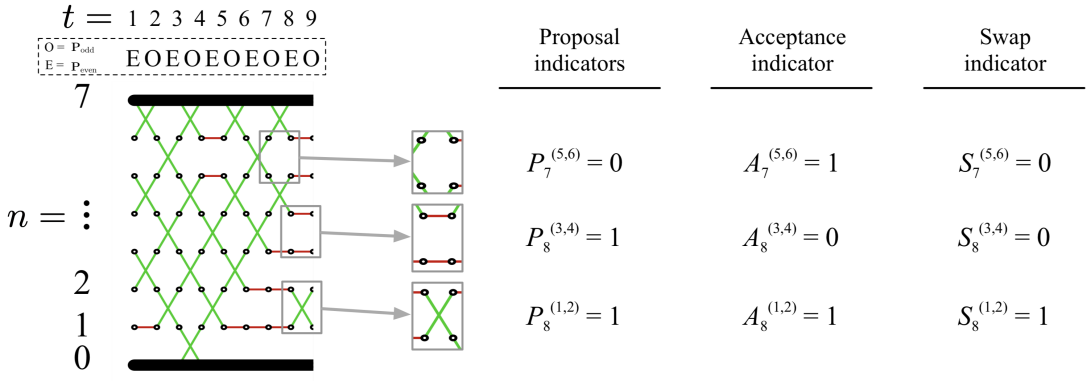

Figure 2.2: Illustration of the proposal, acceptance and swap indicators.

machine m.

In general the replica process for a given machine m is not Markovian, but $(\mathbf{Y}_t, \mathbf{I}_t, \boldsymbol{\epsilon}_t)$ is and provides an alternative but equivalent view of the PT Markov chain. We let $P_t^{(n,n+1)} \in \{0,1\}$ denote an indicator that a swap is proposed between chains n and $n+1$ at scan t. The realized swaps are then defined from the proposal indicators as $S_t^{(n,n+1)} = P_t^{(n,n+1)} A_t^{(n,n+1)}$, where $A_t^{(n,n+1)} | \mathbf{X}_t \sim \text{Bern}(\alpha^{(n,n+1)}(\mathbf{X}_t))$ are acceptance indicator variables (Figure 2.2). The proposal indicators define the communication kernel from the previous section. Equipped with this notation, we can show that for each $m = 0, \ldots, N$, the replica process $(Y_t^m, I_t^m, \epsilon_t^m)$ for machine m satisfies the following recursive relation: initialize $I_0^m = m$ and $\epsilon_0^m = 1$ if $P_0^{(m,m+1)} = 1$ and -1 otherwise. For $t > 0$, we have

$$
\begin{aligned}
Y_{t+1}^m &\sim K_{\beta_{I_t^m}}(Y_t^m, \cdot), \\
I_{t+1}^m &= \begin{cases} I_t^m + \epsilon_t^m & \text{if } S_t^{(I_t^m, I_t^m + \epsilon_t^m)} = 1, \\ I_t^m & \text{otherwise,} \end{cases}, \\
\epsilon_{t+1}^m &= \begin{cases} 1 & \text{if } P_t^{(I_{t+1}^m, I_{t+1}^m + 1)} = 1. \\ -1 & \text{otherwise.} \end{cases}.
\end{aligned}
$$

This is summarized in Algorithm 2. Note that for a fixed m, the replica process $(Y_t^m, I_t^m, \epsilon_t^m)$ is not Markovian since different machines must communicate through the swaps.

Algorithm 2 DistributedDEO

Input: Initial state $\boldsymbol{x}_0$, annealing path π, annealing schedule $\mathcal{B}_N$, number of scans t_{scan}

▷ Initialize chain

$\boldsymbol{x} \leftarrow \boldsymbol{x}_0$

▷ For a machine m, $\mathbf{n}^m$ gives the annealing parameter index for that machine. Initialized with the identity map.

$\boldsymbol{n} \leftarrow (0, 1, 2, \ldots, N)$

▷ Initialize index process. For an annealing parameter index n, $\boldsymbol{i}^n$ gives the machine processing the corresponding chain. This is the inverse of the above mapping.

$\boldsymbol{i} \leftarrow (0, 1, 2, \ldots, N)$

for t **in** 1, 2, …, t_{scan} **do**

▷ Non-reversibility inducing alternation

  if t is even **then**

    $P \leftarrow \{n : 0 \leq n < N, n \text{ is even}\}$

  else

    $P \leftarrow \{n : 0 \leq n < N, n \text{ is odd}\}$

  end if

  ▷ Distributed local exploration phase

  for m **in** $0, \ldots, N$ **do**

  ▷ Fetch current annealing parameter for machine m

    $\beta \leftarrow \beta_{\boldsymbol{n}^m}$

    $x^m \sim K_\beta(x^m_{t-1}, \cdot)$

  end for

  for n **in** $0, \ldots, N-1$ **do**

  ▷ Distributed communication phase

    $\alpha \leftarrow \alpha^{(n,n+1)}$

    ▷ Compute acceptance probability using Equation (2.3)

    $A \sim \text{Bern}(\alpha)$

    if $n \in P$ **and** $A = 1$ **then**

    ▷ Machines swap annealing parameters indices

      $(\boldsymbol{i}^n, \boldsymbol{i}^{n+1}) \leftarrow (\boldsymbol{i}^{n+1}, \boldsymbol{i}^n)$

    end if

  end for

  ▷ Recompute $\mathbf{n}$ from $\boldsymbol{i}$ such that $\boldsymbol{i}^n = m \iff \mathbf{n}^m = n$

  for n **in** $0, \ldots, N$ **do**

    $m \leftarrow \boldsymbol{i}^n$

    $\mathbf{n}^m \leftarrow n$

  end for

  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}$

  $\boldsymbol{i}_t \leftarrow \boldsymbol{i}$

end for

return $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t_{\text{scan}}}), (\boldsymbol{i}_0, \ldots, \boldsymbol{i}_{t_{\text{scan}}})$

2.3.1 Index process

This distributed view of parallel tempering is useful to analyze the efficiency of the algorithm. We can study how efficiently the reference and target can communicate, by measuring how quickly a state traverses from the reference to the target on each machine m. Recall from Section 1.2.1, that I_t^m for $t = 1, 2, \dots$ is the sequence of the indices of the annealing parameters stored on machine m, and encodes the flow of information between the reference and target on machine m. We will refer to the *index process* for machine m as (I_t^m, ϵ_t^m). Refer to the bold paths illustrating a single index process for $N = 8$, $N = 30$ in Figure 1.4. We will use the dynamics of the index process to explain the differences between SEO and DEO communication. The only difference between the two is in the proposal indicators. Define $\mathbf{P}_t = (P_t^{(0,1)}, P_t^{(1,2)}, \dots, P_t^{(N-1,N)})$, $\mathbf{P}_t$ is deterministic for DEO, i.e. $\mathbf{P}_t = \mathbf{P}_{\text{even}} = (1, 0, 1, \dots)$ for even t and $\mathbf{P}_t = \mathbf{P}_{\text{odd}} = (0, 1, 0, \dots)$ for odd t. In SEO, we have $\mathbf{P}_t \sim \text{Unif}\{\mathbf{P}_{\text{even}}, \mathbf{P}_{\text{odd}}\}$.

For SEO, the variables $\epsilon_t^m \sim \text{Unif}\{-1, 1\}$ are i.i.d. for a fixed m, and consequently the index process exhibits a random walk behaviour on each machine. In contrast, for DEO, we have $\epsilon_{t+1}^m = \epsilon_t^m$ so long as $I_{t+1}^m = I_t^m + \epsilon_t^m$ and $\epsilon_{t+1}^m = -\epsilon_t^m$ otherwise. Therefore the index processes for DEO persist in direction as ϵ_{t+1} is only reversed when a swap involving machine m is rejected or if the boundary is reached. This means DEO facilitates a more systematic communication between the reference and target. The qualitative differences between the two regimes can be seen in Figure 1.4, and in Figure 2.3. In particular the index process for DEO in these figures behaves very differently as N increases. We will explore this relation formally in Chapter 5 by formally deriving the scaling limit for the index process.

As mentioned in the introduction, we refer to the PT algorithm with SEO and DEO communication as *reversible* PT and *non-reversible* PT respectively. Our terminology is somewhat abusive but is justified by the analysis in Section 3.1.3 and Section 5.2, where it is shown that, under certain assumptions, the index process is reversible for SEO while it is non-reversible for DEO.

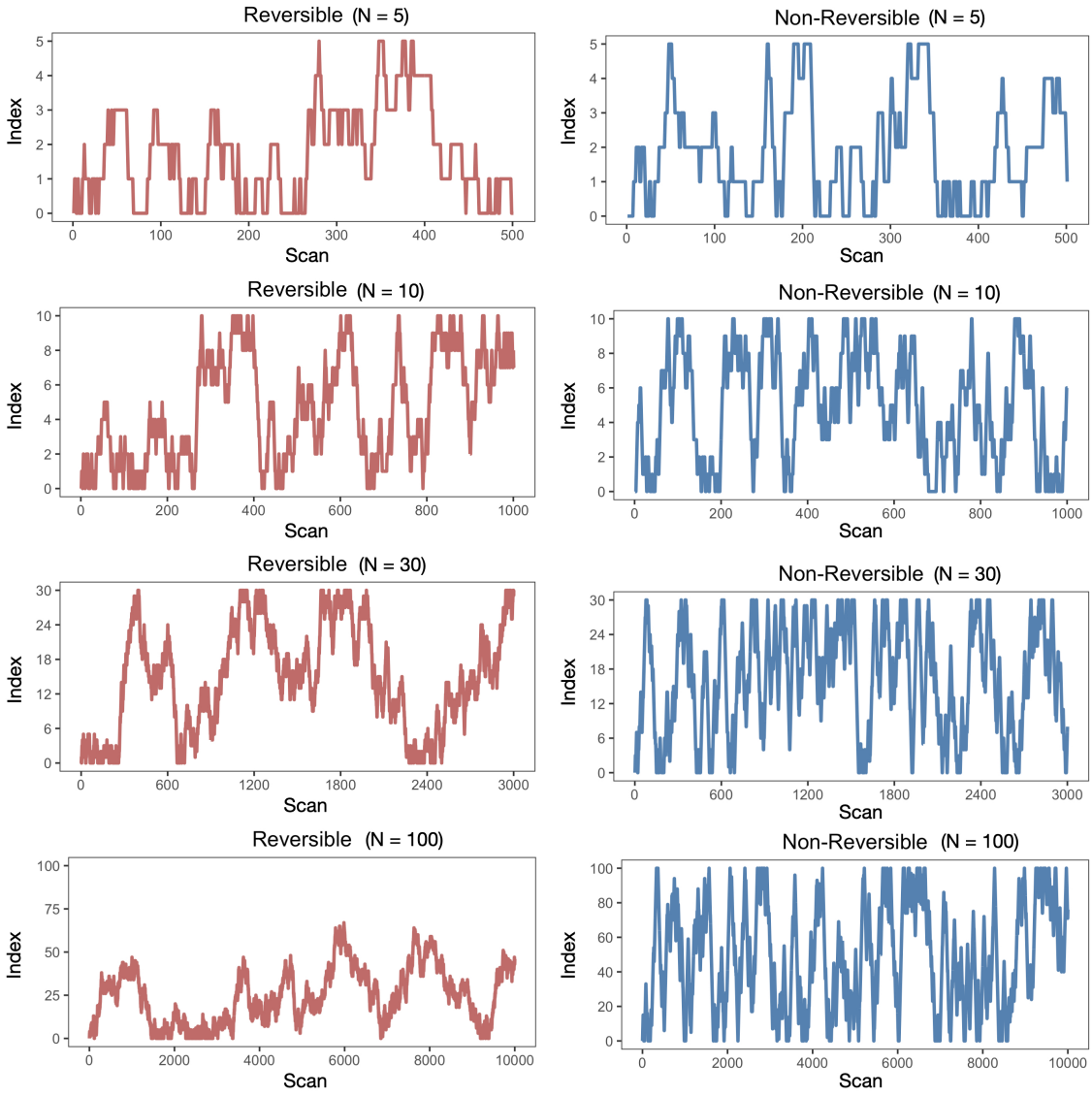

Figure 2.3: Sample trajectories of the index process for machine $m = 0$ for a Gaussian model with $\Lambda = 5$ with the optimal schedule derived in Section 3.3 for $N = 5, 10, 30, 100$. The trajectories are run over $100N$ scans for reversible PT (left) and non-reversible PT (right).

2.4 Performance metrics for PT methods

2.4.1 Effective sample size

The performance of PT is sensitive to both the local exploration and communication moves. The quantity commonly used to evaluate the performance of MCMC algorithms is the effective sample size (ESS); however, ESS measures the combined performance of local exploration and communication, and is not able to distinguish between the two. Since the major difference between PT and standard

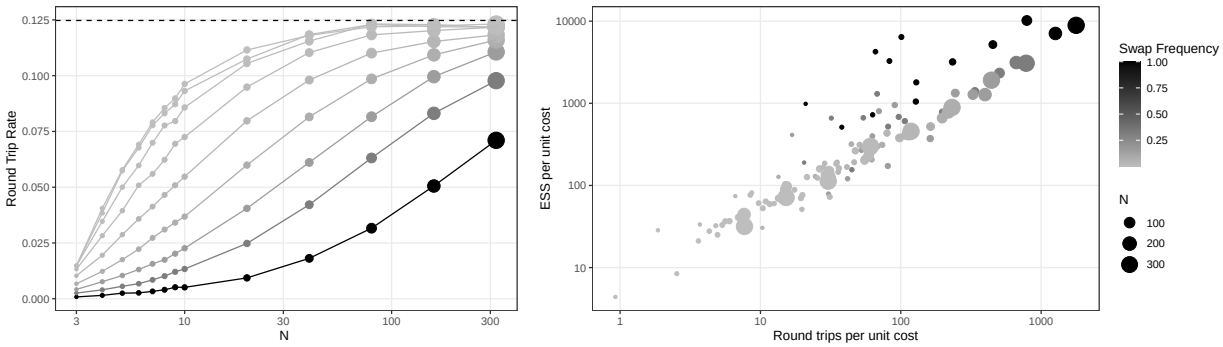

Figure 2.4: The round trip rate and ESS for a 5×5 Ising Model with a magnetic moment 0.1 and t_{expl} 1-bit flips between scans ranging from 1 to 400, and N ranging from 3 to 320. The schedule is tuned using Algorithm 5 with $t_{\text{tune}} = 25000$ scans for tuning, $t_{\text{sample}} = 25000$ scans for sampling. (Left) Round trip rate versus N for different swap attempt frequency ($1/t_{\text{expl}}$). The dotted line is the optimal round trip rate $\bar{\tau}$ predicted by Theorem 6. (Right) The ESS per unit cost versus round trip rate per unit cost for each run, with a correlation coefficient of 0.81.

MCMC is the presence of a communication step, we require a way to measure communication performance in isolation such that we can compare PT methods without dependence on the details of the problem specific local exploration move. We will show that the *round trip rate* is an alternative performance measure from the PT literature (Katzgraber et al., 2006; Lingenheil et al., 2009) that is designed to assess communication efficiency alone.

2.4.2 Round trip rate

We are motivated by the Bayesian context where it is typically possible to obtain one independent sample from the reference distribution π_0 (i.e. from the prior distribution) at each iteration. We say that an *annealed restart* has occurred on machine m when I_t^m goes from 0 to N (i.e. β goes from 0 to 1), which corresponds to a sample generated from π_0 propagating to the target π_1. Informally an annealed restart can be thought of as a sampling equivalent to what is known in optimization as a random restart. We say a *round trip* has occurred on machine m when I_t^m goes from 0 to N and then goes back to 0 (i.e. β goes from 0 to 1 to 0).

Formally, we recursively define $T_{\downarrow,0}^m = \inf\{t : (I_t^m, \epsilon_t^m) = (0, -1)\}$ and for $i \geq 1$,

$$T_{\uparrow,i}^m = \inf\{n > T_{\downarrow,i-1}^m : (I_t^m, \epsilon_t^m) = (N, 1)\},$$
$$T_{\downarrow,i}^m = \inf\{n > T_{\uparrow,i}^m : (I_t^m, \epsilon_t^m) = (0, -1)\}.$$

The i-th annealed restart and round trip for machine m occurs at scan $T^m_{\uparrow,i}$ and $T^m_{\downarrow,i}$ respectively. Let $\mathcal{T}_t$ and $\mathcal{R}_t$ be the total number of annealed restarts and round trips respectively during the first t scans of PT.

We wish to optimize for the percentage of iterations that result in an annealed restart, i.e. $\tau := \lim_{t\to\infty} \mathbb{E}[\mathcal{T}_t]/t$, where we use abusively the same random variables for SEO and DEO but differentiate these schemes by using the probability measures $\mathbb{P}_{\text{SEO}}$ and $\mathbb{P}_{\text{DEO}}$ with associated expectation operators $\mathbb{E}_{\text{SEO}}$ and $\mathbb{E}_{\text{DEO}}$. We use $\mathbb{P}$ and $\mathbb{E}$ for statements that hold for both algorithms. If $\mathcal{T}^m$ and $\mathcal{R}^m$ are the total number of annealed restarts and round trips during the first t scans on machine m respectively, then we have

$$\mathcal{R}^m_t \leq \mathcal{T}^m_t \leq \mathcal{R}^m_t + 1. \tag{2.5}$$

Consequently, by summing (2.5) from $m = 0, \ldots, N$, we have

$$\mathcal{R}_t \leq \mathcal{T}_t \leq \mathcal{R}_t + N + 1.$$

By taking limit as $t \to \infty$ and using the squeeze theorem we have

$$\tau = \lim_{t\to\infty} \frac{\mathbb{E}[\mathcal{R}_t]}{t}.$$

In the PT literature, τ is commonly referred to as the *round trip rate* and has been used to compare the effectiveness of various PT algorithms (Katzgraber et al., 2006; Lingenheil et al., 2009; Jacka and Hernández-Hernández, 2019). Empirically we observed that round trips per unit cost strongly correlate with ESS per unit cost as seen in Figure 2.4, making the round trip rate a natural objective function to compare and tune parallel tempering algorithms.

2.4.3 Expected square jump distance

Another performance metric commonly used in the PT literature is the *expected square jump distance* (ESJD) (Atchadé et al., 2011; Kone and Kofke, 2005). While this criterion is useful within the context of reversible PT for selecting the optimal number of parallel chains, the ESJD is too coarse to compare reversible to non-reversible PT methods as, for any given annealing schedule, the ESJD

is identical in both cases. Moreover, it is not well motivated for non-reversible PT since the annealing parameter process does not exhibit diffusive behaviour (see Figure 1.4 and Section 5.2).

Chapter 3

Non-reversible parallel tempering

The hammers must be swung in cadence, when more than one is hammering the iron.

— Giordano Bruno

In this chapter, we will analyze the non-asymptotic and asymptotic performance for both reversible and non-reversible PT with the assumption that π is the linear annealing path between reference π_0 and target π_1 constructed in Section 2.1.1. We will expand this theory to more general annealing paths in Chapter 4.

3.1 Non-asymptotic analysis of PT algorithms

3.1.1 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 or distributed computational setup and sampling once from each of the local exploration kernel K_β has cost $O(t_{\text{expl}})$ which dominates the cost of the swap kernel $\mathbf{K}^{(n-1,n)}$. Consequently a scan of PT has cost $O(t_{\text{expl}})$ and for a fixed computational budget t_{total}, the total number of scans is $t_{\text{scan}} = O(t_{\text{total}}/t_{\text{expl}})$.

The assumption that the per-iteration cost of PT is independent of the number of chains is reasonable in GPU and parallel computing scenarios, since the communication cost for each swap does not increase with the dimension of the problem (by swapping annealing parameters instead of states). We also assume that the number of PT scans will dominate the number of parallel cores available, i.e. $t_{\text{scan}} \gg N$. This is reasonable when addressing challenging sampling problems. Although there are numerous empirical studies on multi-core and distributed implementation of PT (Altekar et al., 2004; Mingas and Bouganis, 2012; Fang et al., 2014), we are not aware of previous theoretical work investigating such a computational model.

3.1.2 Model assumptions

The analysis of the round trip times is in general intractable because the index process is not Markovian. Indeed, simulating a transition depends on the swap indicators from Section 2.3). Recall a swap indicator for a proposed swap move is $S_t^{(n-1,n)} \sim \text{Bern}(\alpha^{(n-1,n)}(\mathbf{X}))$ where the acceptance probability equals

$$\alpha^{(n-1,n)}(\mathbf{X}) = 1 \wedge \exp((\beta_n - \beta_{n-1})(V(X^{n-1}) - V(X^n)). \tag{3.1}$$

Notice this implies the swap indicator depend on the state configuration $\mathbf{X}$ through $V(X^{n-1})$ and $V(X^n)$. To simplify the analysis, we will make in the remainder of the chapter the following simplifying assumptions:

(A1) *Stationarity*: $\mathbf{X}_0 \sim \boldsymbol{\pi}$ and thus $\mathbf{X}_t \sim \boldsymbol{\pi}$ for all t as the kernel $\mathbf{K}_t^{\text{PT}}$ is $\boldsymbol{\pi}$-invariant.

(A2) *Efficient Local Exploration (ELE)*: For $\mathbf{X} \sim \boldsymbol{\pi}$ and $\bar{\mathbf{X}}|\mathbf{X} \sim \mathbf{K}^{\text{expl}}(\mathbf{X}, \mathrm{d}\bar{\boldsymbol{x}})$, the random variables $V(X^n)$ and $V(\bar{X}^n)$ are independent for all $n = 1, \ldots, N$.

It follows from Assumptions (A1)–(A2) and (2.3) that the behaviour of the communication scheme only depends on the distribution of the state $\mathbf{X}_t$ via the $N+1$ univariate distributions of the chain-specific log-likelihood $V(X^n)$, $n \in \{0, 1, 2, \ldots, N\}$. This allows us to build a theoretical analysis which makes no structural assumption on the state space $\mathcal{X}$ or the target π_1 as typically done in the literature: for example, Atchadé et al. (2011) assume a product space $\mathcal{X} = \mathcal{X}^d$ for large d, and Predescu et al. (2004) assume π_β satisfies a constant heat capacity (i.e. $\beta^{-2}\text{Var}_\beta[V]$ is constant).

Admittedly the ELE assumption (A2) does not hold in practical applications. ELE can be approximated by increasing the number of local exploration kernels applied between consecutive swap (t_{expl}). However one may worry that to achieve a good approximation in challenging problems, t_{expl} would have to be set to a value so large as to defy the practicality of our analysis. Surprisingly, we have observed empirically that this was not the case in the multimodal problems we considered. Figure 3.1 displays results in four models where a local exploration kernel alone induces good mixing of the energy chain $V(X_t^N)$ (hence ELE can be approximated) yet the local exploration kernel alone is insufficient to achieve good mixing on the full state space, $\mathbf{X}_t$ (so that PT is justified and

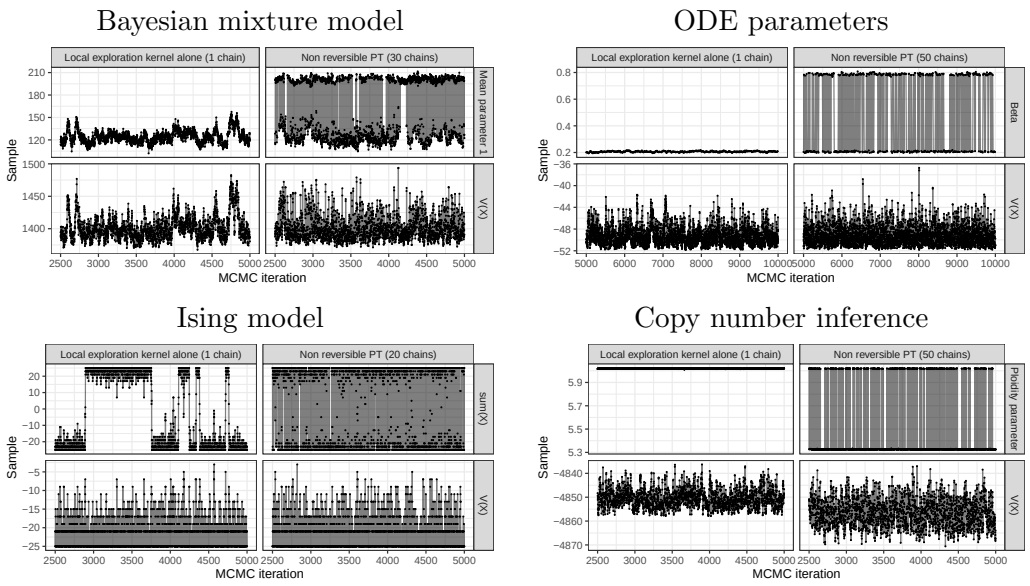

Figure 3.1: Four multimodal examples (described in Section 3.4.1) where a local exploration kernel provides a reasonable approximation of the ELE assumption. For each inference problem, we show trace plots for MCMC based on single chain (i.e. the local exploration kernel alone; left facet), and for a non-reversible PT algorithm based on the same local exploration kernel (right facet). The top facets each show a component of chain X_t^N, and the bottom facets, the energy $-V(X_t^N)$ for the chain targeting π_1.

indeed yields efficient exploration of the configuration space). This gap is possible since $V(X)$ is 1-dimensional and potentially unimodal even when X is not. This is the motivation for ELE since the independence of $V(X)$ and $V(\bar{X})$ is weaker than assuming the independence of X and $\bar{X}$ (as hypothesized e.g. in Section 5.1 of Atchadé et al. (2011)) defeating the need for PT in the first place. Obviously ELE is still expected to be a somewhat crude simplifying assumption in very complex problems; e.g. for the highly challenging high-dimensional copy number inference problem illustrated in Figure 3.1.

In Section 3.4.2, we describe additional empirical results supporting that ELE is a useful model for the purpose of designing and analyzing PT algorithms. In the severe ELE violation regime, we show that the key quantities used in our analysis are either well approximated, or approached as N increases.

Assumptions (A1)-(A2) allow us to express the swap indicators as *independent* Bernoulli random

variables $S_t^{(n-1,n)} \sim \text{Bern}(s^{(n-1,n)})$ where $s^{(n-1,n)}$ is given by the expectation of Equation (3.1),

$$\begin{aligned} s^{(n-1,n)} &:= \mathbb{E}_{\boldsymbol{\pi}}\left[\alpha^{(n-1,n)}(\mathbf{X})\right] \\ &= \mathbb{E}\left[1 \wedge \exp\left((\beta_n - \beta_{n-1})(V(X^{n-1}) - V(X^n))\right)\right]. \end{aligned}$$

where the second expectation is over two independent random variables $X^n \sim \pi_{\beta_n}$ and $X^{n+1} \sim \pi_{\beta_{n+1}}$.

3.1.3 Reversibility and non-reversibility of the index process

Under assumptions (A1)-(A2), each machine's index process (I_t^m, ϵ_t^m) is Markovian for $m = 0, \ldots, N$ with transition kernel K^{SEO} and K^{DEO} for reversible and non-reversible PT respectively. We drop the superscript m when the particular machine is not relevant.

For PT with SEO communication and index process initialized at (I_0, ϵ_0), we can construct the Markov transition kernel K^{SEO} satisfying $(I_{t+1}, \epsilon_{t+1}) \sim K^{\text{SEO}}((I_t, \epsilon_t), \cdot)$ in two steps. In the first step, simulate $I_{t+1}|(I_t, \epsilon_t)$,

$$I_{t+1}|(I_t, \epsilon_t) \sim \begin{cases} (I_t + \epsilon_t) \wedge N \vee 0 & \text{with probability } s^{(I_t, I_t+\epsilon_t)}, \\ I_t & \text{otherwise,} \end{cases} \tag{3.2}$$

where the expression "$\wedge N \vee 0$" enforces the annealing parameter boundaries. In the second step, independently sample

$$\epsilon_{t+1} \sim \text{Unif}\{-1, +1\}.$$

Similarly for DEO, initialize the index process at (I_0, ϵ_0). Analogous to the SEO construction, we construct K^{DEO} satisfying $(I_{t+1}, \epsilon_{t+1}) \sim K^{\text{DEO}}((I_t, \epsilon_t), \cdot)$ in two steps. We first update $I_{t+1}|(I_t, \epsilon_t)$ as in (3.2), but apply the deterministic update in the second step,

$$\epsilon_{t+1} = \begin{cases} \epsilon_t & \text{if } I_{t+1} = I_t + \epsilon_t, \\ -\epsilon_t & \text{otherwise.} \end{cases}$$

The kernel P^{SEO} defines a reversible Markov chain on $\{0, \ldots, N\} \times \{-1, 1\}$ with uniform

stationary distribution while K^{DEO} satisfies the skew-detailed balance condition with respect to the same distribution,

$$K^{\mathrm{DEO}}((i,\epsilon),(i',\epsilon')) = K^{\mathrm{DEO}}((i',-\epsilon'),(i,-\epsilon)),$$

and is thus non-reversible. It falls within the generalized Metropolis–Hastings framework, see, e.g., Lelièvre et al. (2010, Section 2.1.4).

Reversibility necessitates that the Markov chain must be allowed to backtrack its movements. This leads to inefficient exploration of the state space. As a consequence, non-reversibility is typically a favourable property for MCMC chains. A common recipe to design non-reversible sampling algorithms consists of expanding the state space to include a "lifting" parameter that allows for a more systematic exploration of the state space (Chen et al., 1999; Diaconis et al., 2000; Turitsyn et al., 2011; Vucelja, 2016).

The index process (I_t, ϵ_t) for non-reversible PT can be interpreted as a "lifted" version of the index process for reversible PT with lifting parameter ϵ_t. Under DEO communication, I_t travels in the direction ϵ_t and only reverses direction when I_t reaches a boundary or when a swap rejection occurs. This "lifting" construction helps explain the qualitatively different behaviour between reversible and non-reversible PT and will be further explored when identifying the scaling limit of (I_t, ϵ_t) in Section 5.2. The lifted PT of Wu (2017) exploits a similar construction but only one of the $N+1$ index processes is lifted instead of all of them for DEO. A lifted version of simulated tempering has also been proposed by Sakai and Hukushima (2016).

3.1.4 Non-asymptotic domination of non-reversible PT

Assumptions (A1)-(A2) ensure that for each $m = 0, \ldots, N$, $\mathcal{R}_t^m$ is a delayed renewal processes where the renewal event is a round trip occurring at time T_i^m. The corresponding inter-arrival times is $T_i^m = T_{\downarrow,i}^m - T_{\downarrow,i-1}^m$ for $i \geq 1$ and $m = 0, \ldots, N$. In particular, for a fixed m, the $\{T_i^m\}_{i=1}^{\infty}$ are independent and identically distributed. By the key renewal theorem, we have

$$\tau = \sum_{m=0}^{N} \lim_{t\to\infty} \frac{\mathbb{E}[\mathcal{R}_t^m]}{t} = \frac{N+1}{\mathbb{E}[T]}, \tag{3.3}$$

where $T \stackrel{d}{=} T_i^m$.

An analytical expression for $\mathbb{E}[T]$ for reversible PT was first derived by Nadler and Hansmann (2007). We derive an alternative proof for reversible PT and extend it to non-reversible PT in Theorem 1.

Theorem 1. *For any annealing schedule* $\mathcal{B}_N = (\beta_0, \ldots, \beta_N)$,

$$\mathbb{E}_{\text{SEO}}[T] = 2(N+1)N + 2(N+1)\Lambda(\mathcal{B}_N), \tag{3.4}$$

$$\mathbb{E}_{\text{DEO}}[T] = 2(N+1) + 2(N+1)\Lambda(\mathcal{B}_N), \tag{3.5}$$

where $\Lambda(\mathcal{B}_N)$ *equals*

$$\Lambda(\mathcal{B}_N) := \sum_{n=1}^{N} \frac{r^{(n-1,n)}}{1 - r^{(n-1,n)}}, \tag{3.6}$$

and $r^{(n-1,n)} := 1 - s^{(n-1,n)}$ *is the probability of rejecting a swap between chains* $n-1$ *and* n.

The proof can be found in Appendix A.1.1.

Intuitively, Theorem 1 implies $\mathbb{E}[T]$ can be decomposed as the independent influence of communication scheme $\mathbf{K}_t^{\text{comm}}$ and schedule $\mathcal{B}_N$ respectively. When all proposed swaps are accepted (i.e. $\pi = \pi_0$), the index process for reversible PT reduces to a simple random walk on $\{0, \ldots, N\}$, whereas for non-reversible PT, the index processes takes a direct path from 0 to N and back. Therefore, the first term in (3.4) and (3.5) represents the expected time for a round trip to occur in this idealized, rejection-free setting. In particular this suggests that DEO is the optimal communication scheme since it achieves the minimal round trip time of $2N$.

The second term of (3.4) and (3.5) are identical and represent the additional time required to account for rejected swaps under schedule $\mathcal{B}_N$. To understand the intuition behind this term we remark that $\frac{r_n}{1-r_n}$ is the expectation of geometric random variable with failure probability r_n and represents number of failures before a replica moves from β_{n-1} to β_n. We can interpret $\Lambda(\mathcal{B}_N)$ in (3.6) as the total number of rejections per chain required for π_0 to communicate with π_1 through the annealing distributions $\pi_{\beta_1}, \pi_{\beta_2}, \ldots, \pi_{\beta_{N-1}}$.

Motivated by Theorem 1, we will refer to $\Lambda(\mathcal{B}_N)$ as the *non-asymptotic communication barrier*. By applying Theorem 1 to Equation (3.3), we get a non-asymptotic formula for the round trip rate

in terms of $\Lambda(\mathcal{B}_N)$.

Corollary 2. *For any annealing schedule $\mathcal{B}_N$ we have*

$$\tau_{\mathrm{SEO}}(\mathcal{B}_N) = \frac{N+1}{\mathbb{E}_{\mathrm{SEO}}[T]} = \frac{1}{2N + 2\Lambda(\mathcal{B}_N)}, \tag{3.7}$$

$$\tau_{\mathrm{DEO}}(\mathcal{B}_N) = \frac{N+1}{\mathbb{E}_{\mathrm{DEO}}[T]} = \frac{1}{2 + 2\Lambda(\mathcal{B}_N)}. \tag{3.8}$$

Consequently, $\tau_{\mathrm{DEO}}(\mathcal{B}_N) > \tau_{\mathrm{SEO}}(\mathcal{B}_N)$ for $N > 1$.

Note that (3.7) implies that $\tau_{\mathrm{SEO}} \leq \frac{1}{2N}$ regardless of the chosen schedule $\mathcal{B}_N$ and so reversible PT penalizes the user for choosing a large number of chains N. However if N is chosen to be too small, then we should expect the rejection probabilities r_n to be high which will lead to a large $\Lambda(\mathcal{B}_N)$. We will see in the following section that for hard problems, controlling $\Lambda(\mathcal{B}_N)$ requires a potentially large N, rendering reversible PT useless. This makes reversible PT extremely sensitive to the choice of N as well as the schedule.

In contrast, Corollary 2 implies that non-reversible PT dominates reversible PT for any $N > 1$ and any annealing schedule $\mathcal{B}_N$. Moreover, (3.8) does not suffer from the same deterioration in performance for choosing a large N. We will see in Theorem 6 in Section 3.2.4 that the round trip rate actually improves with N and is robust to the choice of schedule.

3.2 Asymptotic analysis of PT algorithms

3.2.1 Rejection rate as divergence

Corollary 2 informs us that the performance of parallel tempering can be studied through $\Lambda(\mathcal{B}_N)$ which depends on the swap statistics. We begin our analysis by studying the dynamics of the swaps between two distributions $p, p' \in \mathcal{P}(\mathcal{X})$ independent of annealing paths. We define the swap and rejection rates $s, r : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, 1]$ respectively as,

$$s(p, p') := \mathbb{E}\left[1 \wedge \frac{p(X')p'(X)}{p(X)p'(X')}\right], \tag{3.9}$$

$$r(p, p') := 1 - s(p, p'), \tag{3.10}$$

where the expectation in (3.9) is taken with respect to independent random variables $X \sim p, X' \sim p'$. The swap rate corresponds to the Metropolis-Hastings acceptance probability for a swap occurring between distributions $p, p' \in \mathcal{P}(\mathcal{X})$. In particular, note that $s^{(n-1,n)} = s(\pi_{\beta_{n-1}}, \pi_{\beta_n})$ for all $n = 1, \ldots, N$. By computing the expectation in (3.9) we can find an alternative representation for the rejection rate,

$$\begin{aligned} r(p, p') &= 1 - \int p(x)p'(x') \wedge p'(x)p(x')\mathrm{d}x\mathrm{d}x' \\ &= \left\| p \times p' - p' \times p \right\|_{\mathrm{TV}}. \end{aligned}$$

This implies that the rejection rate $r(p, p')$ is equal to the total variation distance between the product measures $p \times p'$ and $p' \times p$ and measures the similarity of p and p' in terms of overlap of their densities (see Neklyudov et al. (2018)[Theorem 1] for details). In general, $r(p, p')$ is equal to zero if and only if $p = p'$ and thus defines a symmetric divergence but is not a metric since the triangle inequality does not hold in general. By applying Pinsker's inequality, we obtain a bound for the square rejection rate

$$r(p, p')^2 \leq \mathrm{SKL}(p, p'), \tag{3.11}$$

where $\mathrm{SKL}(p, p')$ is the symmetric Kullback-Leibler (SKL) divergence defined as,

$$\mathrm{SKL}(p, p') := \frac{1}{2}(\mathrm{KL}(p, p') + \mathrm{KL}(p', p)),$$

and $\mathrm{KL}(p, p') := \int_{\mathcal{X}} \log\left(\frac{p(x)}{p'(x)}\right) p(x)\mathrm{d}x$ is the Kullback-Leibler (KL) divergence.

Given $p(x) \propto \exp(W(x))$ and $p'(x) \propto \exp(W'(x)) \in \mathcal{P}(\mathcal{X})$, define $\bar{W}(x) = \frac{1}{2}(W(x) + W'(x))$ and $\bar{p}(x) \propto \exp(\bar{W}(x)) \in \mathcal{P}(\mathcal{X})$ as corresponding to the midpoint of the linear path connecting p, p'. Then it follows from the computation in Predescu et al. (2004, Equation (6)), that we can rewrite the rejection rate in terms of an expectation of $X, X' \overset{\text{i.i.d.}}{\sim} \bar{p}$,

$$r(p, p') = 1 - \frac{\mathbb{E}_{\bar{p}}[\exp(-\frac{1}{2}|\Delta W(X) - \Delta W(X')|]}{\mathbb{E}_{\bar{p}}[\exp(-\frac{1}{2}(\Delta W(X) - \Delta W(X'))]}, \tag{3.12}$$

where $\Delta W(x) = W'(x) - W(x)$ is the log-likelihood ratio between p and p'. By applying a Taylor

expansion to (3.12) we extend Proposition 1 in Predescu et al. (2004) and arrive at the following theorem.

Theorem 3. *If ΔW^3 is integrable with respect to p and p', then $\mathbb{E}_{\bar{p}}[|\Delta W|^3] < \infty$ and for some $C > 0$ independent of p, p'*

$$\left| r(p,p') - \frac{1}{2}\mathbb{E}_{\bar{p}}\left[|\Delta W(X) - \Delta W(X')|\right] \right| \leq C\mathbb{E}_{\bar{p}}[|\Delta W|^3],$$

where $X, X' \overset{i.i.d.}{\sim} \bar{p}$.

See Appendix A.1.2 for the proof.

A subtle but important point to make is that the rejection rate $r(p,p')$ is defined independently of the path connecting p and p', however Theorem 3 suggests that the rejection rate $r(p,p')$ is implicitly interpolating along a linear annealing path between p, p'. This motivates a deep connection of parallel tempering to the linear path, which we will see in the next section and also gives insights into how we can expand the PT framework to non-linear paths, which we will explore in Chapter 4.

3.2.2 The local communication barrier

Now we return out attention to the linear path π_β interpolating between π_0 and π_1. We define the swap and rejection rate $s, r : [0,1]^2 \to [0,1]$ respectively by $s(\beta,\beta') = s(\pi_\beta, \pi_{\beta'})$ and $r(\beta,\beta') = r(\pi_\beta, \pi_{\beta'})$. In this case (3.9) and (3.10) simplify to an expectation with respect to independent one-dimensional random variables $V(X), V(X')$,

$$\begin{aligned} s(\beta,\beta') &= \mathbb{E}\left[1 \wedge \exp((\beta'-\beta)(V(X) - V(X'))\right], \\ r(\beta,\beta') &= 1 - s(\beta,\beta'), \end{aligned}$$

where $X \sim \pi_\beta$ and $X' \sim \pi_{\beta'}$ are independent.

To study the dynamics of the parallel tempering as $N \to \infty$ or equivalent as $\|\mathcal{B}_N\| \to 0$, we need to study the behaviour of $r(\beta,\beta')$ when $\beta \approx \beta'$. The key quantity that drives this asymptotic

regime is given by a function $\lambda : [0,1] \to [0,\infty)$

$$\lambda(\beta) := \lim_{\Delta\beta \to 0} \frac{r(\beta, \beta + \Delta\beta)}{|\Delta\beta|}.$$

Using Theorem 3, we can show that the rejection rate is approximated up to second order by λ.

Theorem 4. *Suppose V^3 is integrable with respect to π_0 and π_1 and $0 \le \beta < \beta' \le 1$, then for some $C > 0$ independent of π_β*

$$\left| r(\beta, \beta') - \int_\beta^{\beta'} \lambda(u) \mathrm{d}u \right| \le C \sup_{u \in [\beta, \beta']} \mathbb{E}_u[|V|^3] |\beta' - \beta|^3,$$

where λ is twice continuously differentiable and equal to

$$\lambda(\beta) = \frac{1}{2} \mathbb{E}_\beta[|V(X) - V(X')|], \qquad X, X' \overset{i.i.d.}{\sim} \pi_\beta. \tag{3.13}$$

See Appendix A.1.3 for the proof.

We can interpret λ as the instantaneous rate of rejection of a proposed swap at annealing parameter β. When $\lambda(\beta)$ is high, swaps are much more likely to be rejected, implying $\lambda(\beta)$ measures how sensitive π_β to small perturbations in β. Motivated by Theorem 4 we will henceforth refer to λ as the *local communication barrier*, since a large $\lambda(\beta)$ makes communicating with a chain at annealing parameter β difficult. See Figure 3.7 (center) for examples of estimated λ from various models.

3.2.3 The global communication barrier

By summing the rejection rate for any annealing schedule $\mathcal{B}_N$ Theorem 4 tells us that the sum of the rejection rates is approximately constant and equal to $\Lambda := \int_0^1 \lambda(\beta) \mathrm{d}\beta$ up to an $O(1/N^2)$ error.

Corollary 5. *Suppose V^3 is integrable with respect to π_0 and π_1, and $\mathcal{B}_N$ is generated by γ, then*

$$\left| \sum_{n=1}^N r(\beta_{n-1}, \beta_n) - \Lambda \right| \le \frac{C \|V^3\|_\pi \|\dot{\gamma}\|_\infty^2}{N^2},$$

where $\|V^3\|_\pi = \sup_{\beta \in [0,1]} \mathbb{E}_\beta[|V|^3] < \infty$.

See Appendix A.1.4 for the proof.

Corollary 5 tells us that when N is large, the sum of the rejection rates is approximately invariant to *any* choice of $\mathcal{B}_N$. This approximate invariance is a surprising and non-trivial fact that will have significant consequences for both the theoretical and methodological advancements of the PT algorithm, which we will explore in Sections 3.2.4 and 3.3.

Since λ is a measure of much π_β deforms for small changes in β, we can interpret Λ as total deformation occuring between π_0 and π_1 along the path π_β. Motivated by Corollary 5 we will refer to Λ as the *global communication barrier*.

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 (3.13) that $\lambda = 0$ if and only if V is constant π_β-a.s. for all $\beta \in [0,1]$ which happens precisely when $\pi_0 = \pi$. So Λ defines a natural symmetric divergence measuring the difficulty of communication between π_0 and π. It can be interpreted as the "total rejection" along the linear path π_β between π_0 and π_1.

3.2.4 Asymptotic domination of non-reversible PT

In the previous section, we performed an asymptotic swap analysis and discovered the communication barrier. We will now use the communication barrier to analyze the asymptotic performance of parallel tempering when the number of parallel chains N is large. Recall from Corollary 2 in Section 3.1.4 that the round trip rate is controlled by the non-asymptotic communication barrier $\Lambda(\mathcal{B}_N)$.

By Corollary 5 the non-asymptotic communication barrier is bounded below by Λ up to an $O(1/N^2)$ error,

$$\Lambda(\mathcal{B}_N) = \sum_{n=1}^{N} \frac{r(\beta_{n-1}, \beta_n)}{1 - r(\beta_{n-1}, \beta_n)} \geq \sum_{n=1}^{N} r(\beta_{n-1}, \beta_n) = \Lambda + O\left(\frac{1}{N^2}\right).$$

This combined with Corollary 2 implies as $N \to \infty$, round trip rate for non-reversible PT satisfies

$$\limsup_{N \to \infty} \tau_{\mathrm{DEO}}(\mathcal{B}_N) \leq \tau_\infty,$$

where $\tau_\infty := (2 + 2\Lambda)^{-1}$ as $N \to \infty$. Theorem 6 shows that as $N \to \infty$, the limit of the round trip rate for both reversible and non-reversible PT exists, and converges to an upper bound as $N \to \infty$

unlike reversible PT which decays to zero. This can be empirically observed in Figure 2.4 (left).

Theorem 6. *Suppose V^3 is integrable with respect to π_0 and π_1 and $\mathcal{B}_N$ is generated by a schedule generator γ. Then as $N \to \infty$ we have:*

(a) The non-asymptotic communication barrier $\Lambda(\mathcal{B}_N)$ converges to Λ and satisfies,

$$|\Lambda(\mathcal{B}_N) - \Lambda| \leq \frac{\|V\|_\pi \|\dot{\gamma}\|_\infty}{N} + O\left(\frac{1}{N^2}\right),$$

where $\|V\|_\pi = \sup_\beta \mathbb{E}_\beta[|V|]$.

(b) The round trip rate for reversible PT, τ_{SEO}, goes to zero:

$$\tau_{\mathrm{SEO}}(\mathcal{B}_N) \sim \frac{1}{2N + 2\Lambda} \longrightarrow 0.$$

(c) The round trip rate for non-reversible PT, τ_{DEO} satisfies

$$\tau_{\mathrm{DEO}}(\mathcal{B}_N) \longrightarrow \tau_\infty > 0,$$

where $\tau_\infty = (2 + 2\Lambda)^{-1}$, and the convergence occurs with the same rate as $\Lambda(\mathcal{B}_N)$.

See Appendix A.1.5 for the proof.

We remark that when N is large, the round trip rate for reversible PT decays to zero regardless of the choice of schedule and is very sensitive to the choice of schedule for low values of N. In contrast, $\tau_{\mathrm{DEO}}(\mathcal{B}_N)$ converges to τ_∞ for any schedule $\mathcal{B}_N$ as long as $\|\mathcal{B}_N\| \to 0$. This shows that when N is large, τ_{DEO} is robust to the choice of schedule. We can see this behaviour in Figure 3.2 (right) comparing reversible and non-reversible PT for both a uniform schedule generated by $\gamma(w) = w$ and the (approximately) optimal schedule derived in Section 3.3.1. We will explore how to overcome this limitation in Chapter 4.

In general, Λ is large when π_0 deviates significantly from π_1 and defines the problem's difficulty. Since Λ is problem-specific, this identifies a limitation of PT present even in its non-reversible flavour. Adding more cores to the task will never be harmful but does have a diminishing return. The bound $\tau_\infty = (2 + 2\Lambda)^{-1}$ could indeed be very small for complex problems. Moreover, it is

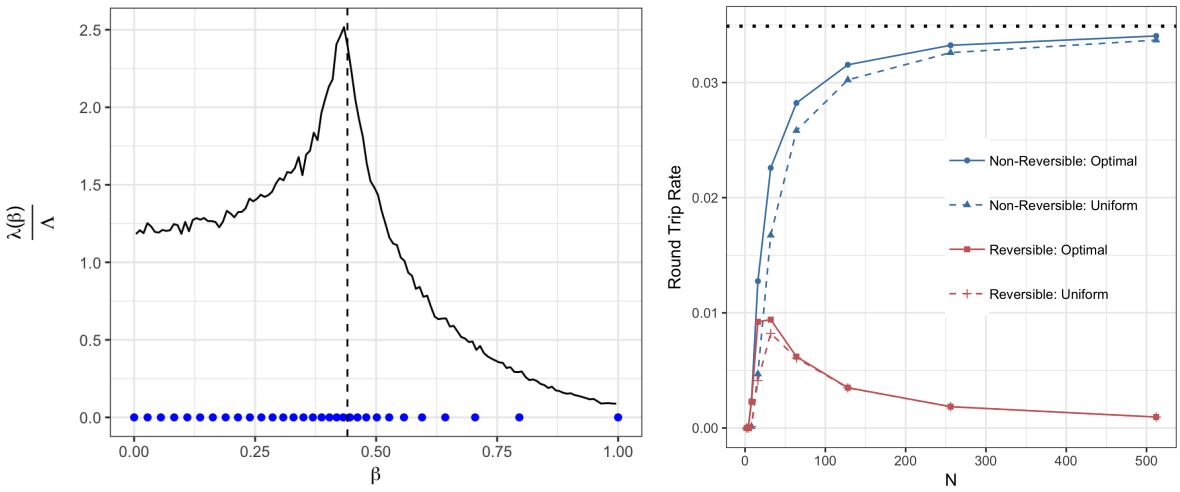

Figure 3.2: (Left) Optimal annealing schedule for the Ising model with $M = 20$, $\Lambda = 13.33$ with $N = 30$ (see Section 3.2.6 for details). The vertical line is at the critical inverse-temperature β_c. (Right) The round trip rates for the Ising model with $M = 20$ as a function of N with uniform schedule (dashed) and the optimal schedule (solid) for both non-reversible PT (blue) and reversible PT (red). The dotted horizontal line represents the approximation of the optimal round trip rate $\hat{\tau}_\infty = (2 + 2\hat{\Lambda})^{-1}$.

independent of the choice of annealing schedule, and hence it cannot be improved by the schedule optimization procedure described in Section 3.3.

3.2.5 High-dimensional scaling of communication barrier

We notice that the asymptotic performance of parallel tempering is entirely controlled by Λ. It is known that parallel tempering suffers from the curse of dimensionality (Atchadé et al., 2011), which should be accounted for by Λ. We determine here the asymptotic behaviour of λ and Λ when the dimension d of $\mathcal{X}$ is large. To make the analysis tractable, we assume that the reference and target distributions factorize as d identical copies of π_0 and π_1 respectively as in Atchadé et al. (2011); Roberts and Rosenthal (2014),

$$\pi_0^{(d)}(x^{(d)}) = \prod_{i=1}^{d} \pi_0(x_i), \qquad \pi_1^{(d)}(x^{(d)}) = \prod_{i=1}^{d} \pi_1(x_i),$$

where $x^{(d)} = (x_1, \ldots, x_d) \in \mathcal{X}^d$. This provides a model for weakly dependent high-dimensional distributions. We only make this structural assumption on the state space and distribution to establish Proposition 7 below.

The corresponding linear annealing path is given by

$$\pi_\beta^{(d)}(x^{(d)}) = \prod_{i=1}^{d} \pi_\beta^{(d)}(x_i). \tag{3.14}$$

Let $\lambda^{(d)}$ and $\Lambda^{(d)}$ be the local and global communication barriers for $\pi_\beta^{(d)}$ respectively. It follows from Proposition 7 that $\lambda^{(d)}$ and $\Lambda^{(d)}$ increase at a $O(d^{1/2})$ rate as $d \to \infty$.

Proposition 7 (High Dimensional Scaling)**.** *If V^3 is integrable with respect to π_0 and π_1, then $d \to \infty$,*

$$\lambda^{(d)}(\beta) \sim \sqrt{\frac{d}{\hat{\pi}} I(\beta)}, \qquad \Lambda^{(d)} \sim \int_0^1 \sqrt{\frac{d}{\hat{\pi}} I(\beta)}\, \mathrm{d}\beta,$$

where $I(\beta) = \mathrm{Var}_\beta[V]$ is the Fisher information of π_β, and $\hat{\pi}$ in the denominator is the constant $3.141\ldots$.

See Appendix A.1.6 for the proof.

We remark $dI(\beta)$ is the Fisher information for the annealing path $\pi_\beta^{(d)}$. In particular, Proposition 7 implies that in the high-dimensional limit, λ^2 scales to a constant multiple of the Fisher information. This fact will be exploited when studying the geometric properties of PT in Chapter 4.

Multimodal decomposition of the communication barrier

Since PT is often used to sample from multimodal targets, it is natural to ask how the communication barrier behaves under the presence of modes. Similar to Woodard et al. (2009), we partition $\mathcal{X}$ into the disjoint union $\mathcal{X} = \bigcup_{k=1}^K \mathcal{X}_k$ where $\mathcal{X}_k$ represents region corresponding to the k-th mode of π and the target decomposes as a mixture of its modes $\pi_1(x) = \sum_{k=1}^K p_k \pi_1(x|\mathcal{X}_k)$ where $p_k = \pi(\mathcal{X}_k)$ and $\pi_1(x|\mathcal{X}_k) = p_k^{-1}\pi_1(x)\mathbb{I}_{\mathcal{X}_k}(x)$ are the probability mass and distribution of the k-th mode. If we assume that the reference distribution puts the same relative mass on each mode as the target, $\pi_0(\mathcal{X}_k) = \pi_1(\mathcal{X}_k)$, then π_β decomposes as

$$\pi_\beta(x) \propto \sum_{i=1}^{K} p_k \pi_\beta(x|\mathcal{X}_k).$$

Similarly $V(x) = \sum_{k=1}^{K} V_k(x)\mathbb{I}_{\mathcal{X}_k}(x)$ where $V_k(x) = -\log(\pi_1(x|\mathcal{X}_k)/\pi_0(x|\mathcal{X}_k))$. Define $\lambda_{k,k'}(\beta)$ and $\Lambda_{k,k'}$ as the local and global communication barrier between mode k and k' by

$$\lambda_{k,k'}(\beta) = \frac{1}{2}\mathbb{E}\left[|V_k(X_k) - V_{k'}(X_{k'})|\right], \qquad \Lambda_{k,k'} = \int_0^1 \lambda_{k,k'}(\beta)\mathrm{d}\beta,$$

where $X_k \sim \pi_\beta(x|\mathcal{X}_k)$ and $X_{k'} \sim \pi_\beta(x|\mathcal{X}_{k'})$ are independent. An immediate consequence of (3.13) in Theorem 4 implies the following decomposition of the communication barrier.

Proposition 8 (Multimodal decomposition)**.** *Suppose V^3 is integrable with respect to π_0 and π_1. If $\pi_0(\mathcal{X}_k) = \pi_1(\mathcal{X}_k)$ then,*

$$\lambda(\beta) = \sum_{k=1}^{K}\sum_{k'=1}^{K} p_k p_{k'} \lambda_{k,k'}(\beta), \qquad \Lambda = \sum_{k=1}^{K}\sum_{k'=1}^{K} p_k p_{k'} \Lambda_{k,k'}, \tag{3.15}$$

In particular (3.15) implies Λ is a weighted average over the communication barriers between modes. For the specific case where the modes are symmetric under a change of labels, we have $\Lambda_{k,k'} = \Lambda$ and thus Λ is invariant to K. We remark that the meta-model used in this section is not realistic for practical problems as we do not know apriori the location of the modes. This meta-model is only used to obtain intuition on the multimodal scalability of the method. Section 3.4.1 illustrates the empirical behaviour of the method in several genuinely challenging multimodal problems.

The takeaway is that parallel tempering is stable if the reference overlaps the modes of the target. If the reference distribution fails to capture the target mass, which is often the case for complex problems in practice, one should expect to see severe deterioration in performance. In Chapter 4, we still discuss how to improve the performance of PT when the reference and target distributions are nearly mutually singular.

3.2.6 Examples

Gaussian model

Suppose $\pi_1 \sim N(0, \sigma_1^2\mathbb{I}_d)$, and $\pi_0 \sim N(0, \sigma_0^2\mathbb{I}_d)$ with $\sigma_0 > \sigma_1$. It can be shown that $\pi_\beta \sim N(0, \sigma(\beta)^2\mathbb{I}_d)$ where $\sigma(\beta)^{-2} = (1-\beta)\sigma_0^{-2} + \beta\sigma^{-2}$. Theorem 1 in Predescu et al. (2004) implies the

following closed form expressions for $\lambda(\beta)$ and $\Lambda(\beta)$,

$$\lambda(\beta) = \frac{2^{1-d}(\sigma_1^{-2} - \sigma_0^{-2})}{B\left(\frac{d}{2}, \frac{d}{2}\right)}\sigma(\beta)^2, \qquad \Lambda = \frac{2^{2-d}}{B\left(\frac{d}{2}, \frac{d}{2}\right)} \log\left(\frac{\sigma_0}{\sigma_1}\right), \tag{3.16}$$

where $B(a, b)$ is the beta function. As $d \to \infty$, we have $\Lambda \sim \sqrt{\frac{2d}{\pi}} \log\left(\frac{\sigma_0}{\sigma_1}\right)$, which is consistent with Proposition 7.

Ising model

We now estimate the communication barrier for the 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 that are nearest neighbours on the lattice, the target distribution is annealed by the inverse temperature β and the tempered distributions are given by

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

This is an M^2 dimensional model which undergoes an approximate phase transition as $M \to \infty$ at some critical inverse-temperature β_c. When $\mu = 0$ it is known that $\beta_c = \log(1 + \sqrt{2})/2$ (Baxter, 2007). We consider experiments with $\mu = 0$ and with $\mu = 0.1$, the latter denoted "Ising with magnetic field" and abbreviated "magnetic" in composite figures.

We observe that λ exhibits very different characteristics in this scenario compared to the Gaussian model: it is not monotonic and is maximized at the critical temperature. Consequently, the optimal annealing schedule is denser near the phase transition. We also note from Figure 3.3 that both λ and Λ increase roughly linearly with respect to M, similarly to the conclusion of Proposition 7, however here the conditions of Proposition 7 do not apply. In Section 3.3.1, we will see as N increases, the round trip rate of reversible PT decays to zero and non-reversible PT increases towards $(2 + 2\Lambda)^{-1}$ (see Figure 3.2). This is consistent with Theorem 6.

Discrete-multimodal problem

Consider a discrete state space $\mathcal{X} = \{1, \ldots, 2k\}$, and let $1_{\text{Even}} : \mathcal{X} \to \{0, 1\}$ denote the indicator function for even numbers. Define $\pi_1(x) = \frac{1}{ka_1 + k} a_1^{1_{\text{even}}(x)}$ and $\pi_0(x) = \frac{1}{ka_0 + k} a_0^{1_{\text{even}}(x)}$ with $a_0, a_1 > 0$.

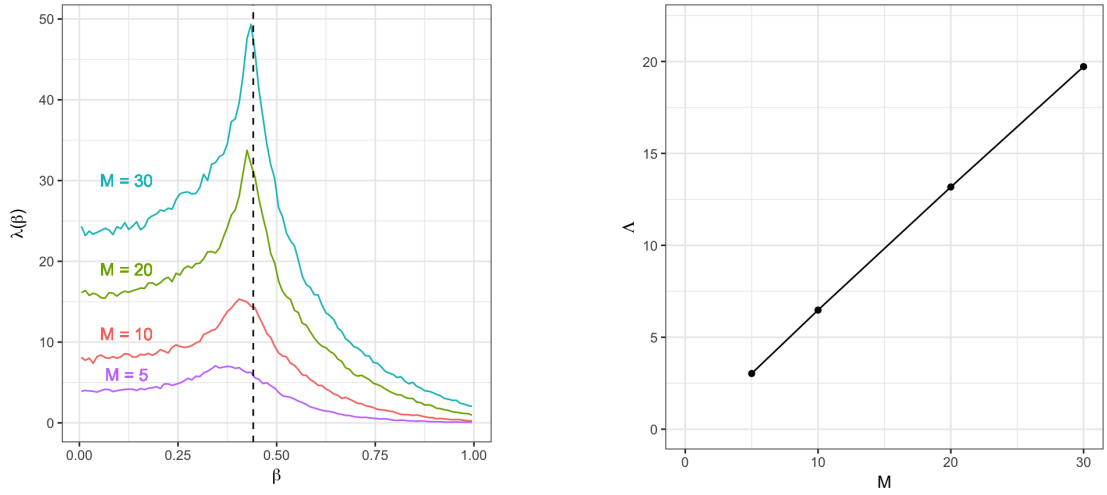

Figure 3.3: Estimate of the local communication barrier (left) and global communication barrier (right) for the Ising model with $\mu = 0$ and $M = 5, 10, 20, 30$. The vertical line is at the phase transition.

In this example a_i represents the relative mass placed on the even sites relative to the odd sites. The log likelihood is $V(x) = 1_{\text{even}}(x) \log \frac{a_1}{a_0}$ with annealing path,

$$\pi_\beta(x) = \frac{1}{Z(\beta)} a_\beta^{1_{\text{even}}(x)},$$

where $a_\beta = a_0^{1-\beta} a_1^\beta$ and $Z(\beta) = k(1 + a_\beta)$. A simple computation using (3.13) shows that the local communication barrier is,

$$\lambda(\beta) = \frac{a_\beta \log(a)}{(1 + a_\beta)^2},$$

where $a = |\frac{a_1}{a_0}|$. By integrating we obtain the global communication barrier between π and π_0,

$$\Lambda = \frac{a - 1}{2(a + 1)}.$$

Notice this is consistent with Proposition 8 and is independent of the number of modes.

3.3 Tuning non-reversible PT

So far we have established that non-reversible PT dominates reversible PT both non-asymptically and asymptotically for any $\mathcal{B}_N$ for a fixed N and as $N \to \infty$. In practice the performance of PT is very sensitive to the choice of schedule. We will establish in this section practical tuning guidelines for non-reversible PT.

3.3.1 Optimal annealing schedule

We first discuss how to optimize the annealing schedule $\mathcal{B}_N$ to maximize τ_{DEO} when N is fixed, which is equivalent to minimizing the non-asymptotic communication barrier inefficiency, $\Lambda(\mathcal{B}_N) = \sum_{n=1}^{N} r_n/(1-r_n)$ where $r_n = r(\beta_{n-1}, \beta_n)$. To get a tractable approximate characterization of the feasible region of $r_1, r_2, \ldots, r_N$, we use Corollary 5, which implies $\sum_{n=1}^{N} r_n \approx \Lambda$ for all schedules $\mathcal{B}_N$. Therefore assuming $\|\mathcal{B}_N\|$ is small enough to ignore the error term, finding the optimal schedule $\mathcal{B}_N^*$ is approximately equivalent to solving the constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{n=1}^{N} \frac{r_n}{1-r_n}, \\ \text{s.t.} \quad & \sum_{n=1}^{N} r_n = \Lambda, \\ & r_n \geq 0. \end{aligned}$$

Using Lagrange multipliers this leads to a solution where the rejection probabilities r_n^* are constant in n, with a shared value denoted by r^*.

Consequently, the optimal schedule $\mathcal{B}_N^* = (\beta_0^*, \ldots, \beta_N^*)$ satisfies $r(\beta_{n-1}^*, \beta_n^*) = r^*$ for all n. Theorem 4 and Corollary 5 imply that r^* must satisfy $r^* \approx \int_{\beta_{n-1}^*}^{\beta_n^*} \lambda(\beta)\mathrm{d}\beta$ for all n and $r^* \approx \Lambda/N$ with an $O(N^{-3})$ error. By equating these two estimates for r^* and summing from $n = 1, \ldots, N$ we get

$$\int_0^{\beta_n^*} \lambda(\beta)\mathrm{d}\beta \approx \frac{n}{N}\Lambda, \quad n = 0, \ldots, N, \tag{3.17}$$

with an error of $O(N^{-2})$. If we ignore error terms, (3.17) implies that $\beta_n^* \approx \gamma(n/N)$ where

Algorithm 3 UpdateSchedule

Input: Communication barrier λ, schedule size N

1: $\Lambda \leftarrow \int_0^1 \lambda(\beta)\mathrm{d}\beta$
2: **for** n in $0, 1, 2, ..., N$ **do**
3: Find β_n^* that solves Equation (3.17) using e.g. bisection.
4: **end for**
5: **return** $\mathcal{B}_N^* = (\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_N^*)$

$\gamma(w) = F^{-1}(w)$ and $F(\beta) = \Lambda(\beta)/\Lambda$ (see Figure 3.4(2)). In general λ is not known but instead estimated from the MCMC output.

The "equi-acceptance" result in (3.17) is consistent with other theoretical frameworks and notions of efficiency (Atchadé et al., 2011; Lingenheil et al., 2009; Kofke, 2002; Predescu et al., 2004). However implementing this equi-acceptance recommendation in practice is non-trivial. Previous work relied on Robbins-Monro schemes (Atchadé et al., 2011; Miasojedow et al., 2013), which introduce sensitive tuning parameters. In contrast, provided the integral of λ can be estimated reliably (which we establish with Algorithm 4 in the next section), (3.17) provides the basis for a straightforward and effective schedule optimization scheme, described in Algorithm 3.

Example: Gaussian

Substituting (3.16) into (3.17) we have that β_n^* satisfies

$$\sigma_{\beta_n^*} = \sigma_0^{1-\frac{n}{N}} \sigma_1^{\frac{n}{N}}.$$

This is the same spacing obtained (based on a different theoretical approach) in Atchadé et al. (2011) and Predescu et al. (2004) for the Gaussian model. This combined with Proposition 8 also justifies why the geometric schedule works for Gaussian mixture models with well-separated modes.

Example: Ising model

The Gaussian model is a rare case where the optimal schedule can be analytically determined. In practice, we need to approximate it using an estimate of the communication barrier. For example, in Figure 3.2 we compute the optimal schedule for the Ising model using Algorithm 3 and using the local communication barrier from Section 3.3. Notice that the optimal schedule is not a geometric

schedule often used in the literature and our algorithm automatically increases the density of annealing parameters near the critical temperature.

3.3.2 Estimation of the communication barrier

The goal of this section is to find an approximation to the communication barrier, or equivalently approximate the *cumulative communication barrier*

$$\Lambda(\beta) := \int_0^\beta \lambda(u)\mathrm{d}u.$$

Assume we have access to a collection of samples $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T)$ from a non-reversible PT scheme based on an arbitrary annealing schedule $\mathcal{B}_N$. These samples may come from a short pilot run, or, as described in the next section, from the previous round of an iterative scheme. For a given schedule $\mathcal{B}_N$, when the central limit theorem for Markov chains holds, the Monte Carlo estimates for the rejection rates satisfy

$$\hat{r}^{(n-1,n)} = \frac{1}{T}\sum_{t=1}^{T} \alpha^{(n-1,n)}(\mathbf{X}_t) = r^{(n-1,n)} + O_p(T^{-1/2}). \tag{3.18}$$

Next, using Theorem 4 we obtain $\sum_{i=1}^{n} r^{(i-1,i)} = \Lambda(\beta_n) + O(N^{-2})$. This motivates the following approximation for $\Lambda(\beta_n)$,

$$\hat{\Lambda}(\beta_n) = \sum_{i=1}^{n} \hat{r}^{(i-1,i)}, \tag{3.19}$$

which has an error of order $O_p(\sqrt{N/T} + N^{-2})$ (see Figure 3.4 (left)).

Given $\hat{\Lambda}(\beta_0), \ldots, \hat{\Lambda}(\beta_N)$, we estimate the function $\hat{\Lambda}(\beta)$ 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 (Fritsch and Carlson, 1980) and denote the monotone interpolation by $\hat{\Lambda}(\beta)$. While we only use $\hat{\Lambda}(\beta)$ in our schedule optimization procedure, it is still useful to estimate $\lambda(\beta)$ for visualization purposes. We use the derivative of our interpolation, $\hat{\lambda}(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\hat{\Lambda}(\beta)$, which is a piecewise quadratic function.

The ideas described in this section so far are summarized in Algorithm 4, which given rejection

Algorithm 4 CommunicationBarrier

Input: Rejection rate $\{r^{(n-1,n)}\}$, schedule $\mathcal{B}_N$

1: For each $\beta_n \in \mathcal{B}_N$, compute $\Lambda(\beta_n)$ using Equation (3.19)
2: $S \leftarrow \{(\beta_0, \Lambda(\beta_0)), (\beta_1, \Lambda(\beta_1)), \ldots, (\beta_N, \Lambda(\beta_N))\}$
3: Compute a monotone increasing interpolation $\Lambda(\cdot)$ of the points S
   ▷ e.g. using Fritsch and Carlson (1980)
4: $\Lambda \leftarrow \Lambda(1)$
5: $\lambda(\beta) \leftarrow \frac{\mathrm{d}}{\mathrm{d}\beta}\Lambda(\beta)$
6: **return** $\lambda(\cdot), \Lambda$

statistics collected for a fixed annealing schedule provides an estimate of the communication barrier. As a byproduct of Algorithm 4 we also obtain a consistent estimator $\hat{\tau} = (2 + 2\hat{\Lambda})^{-1}$ for the optimal round trip rate τ_∞, where $\hat{\Lambda} = \hat{\Lambda}(1)$. We show in Figure 3.8 an example where the ELE assumption (A2) is severely violated, yet λ is still reliably estimated. This allows us to compare the empirically observed round trip rate against $\hat{\tau}$, and hence estimate how far an implementation deviates from optimal performance.

3.3.3 Tuning N

Theorem 6 shows that non-reversible PT does not deteriorate in performance as we increase N unlike reversible PT, however the gains in round trip rate eventually become marginal. When N is very large we expect to accumulate more round trips by running $k > 1$ parallel copies of PT. As we shall see, the large N asymptotic is still useful however in order to determine the optimal number k^* of PT copies.

Suppose there are k copies of PT running in parallel consisting of $N + 1$ chains with optimal annealing schedule $\mathcal{B}_N$. If $\bar{N}$ is the total number of cores available then k and N satisfy the constraint $k(N + 1) = \bar{N}$. By Corollary 2 the total round trip rate across all k copies of PT is

$$\tau = k\tau_{\mathrm{DEO}}(\mathcal{B}_N) = \frac{\bar{N}}{2(N+1)(1+\Lambda(\mathcal{B}_N))}. \tag{3.20}$$

From Section 3.3.1, the optimal schedule $\mathcal{B}_N^*$ has a corresponding swap rejection rate $r^* = \Lambda/N + O(N^{-3})$. Substituting this into (3.20) we get $\tau = \tau_\Lambda(N) + O(N^{-1})$, where

$$\tau_\Lambda(N) = \frac{\bar{N}(1 - \Lambda/N)}{2(N+1)(1-\Lambda/N+\Lambda)}.$$

Ignoring error terms, $\tau_\Lambda(\cdot)$ is optimized when we run $k^* = \bar{N}/(1+N^*)$ copies of PT with $N^*+1 = 2\Lambda+1$ chains to achieve an optimum round trip rate

$$\tau^* = \frac{k^*}{2+4\Lambda} = \frac{\bar{N}}{2(1+2\Lambda)^2}.$$

Optimal rejection rate

Note that when τ is optimized, we have $r^* \approx 1/2$, which differs from the 0.77 optimal rejection rate from the reversible PT literature (Atchadé et al., 2011; Kofke, 2002; Predescu et al., 2004). In Section 3.4.2, we show empirically that when the ELE assumption (A2) is severely violated, the recommendation in the present section is turned into a bound, $r^* < 1/2$, as increasing N appears to alleviate ELE violations.

3.3.4 Iterative schedule optimization

Suppose we have $\bar{N}$ cores available and a computational budget of t_{scan} scans of PT, one scan consisting in one application of $\mathbf{K}^{\text{PT}}$. The first t_{tune} scans are used to find an accurate estimate of the communication barrier $\hat{\lambda}(\cdot)$ and the remaining $t_{\text{sample}} = t_{\text{scan}} - t_{\text{tune}}$ scans to sample from the target. Algorithm 5 approximates $\hat{\Lambda}$ with error $O_p(\sqrt{\bar{N}/t_{\text{tune}}} + \bar{N}^{-2})$ by iteratively using Algorithms 3 and 4 for maxRound$= \log_2 t_{\text{tune}}$ rounds until the tuning budget is depleted. Equipped with $\hat{\Lambda}(\cdot)$, we can use the remaining t_{sample} scans to run $k^* = \bar{N}/(N^*+1)$ copies of non-reversible PT with $N^* = 2\hat{\Lambda}$ chains with optimal schedule $\mathcal{B}^*_{N^*}$ from Algorithm 3. In cases where it is suspected that (A2) may be severely violated (as evidenced for example by a large gap between $\bar{\tau}$ and the estimated τ as described in Section 3.3.2), it may be advantageous to also attempt $N^* = \bar{N}$ in line 10 of Algorithm 5. We show empirically in Figure 3.7 (top) that in a range of synthetic and real world problems our scheme converges using a small number of iterations, in the order of maxRound $= 10$. In our experiments we used $t_{\text{tune}} = t_{\text{scan}}/2$ and discarded the samples from the first t_{tune} scans as burn-in.

3.3.5 Normalizing constant computation

Algorithm 5 can be easily modified to obtain an asymptotically unbiased estimators for quantities of the form $\int_0^1 \mathbb{E}_\beta[f(V(X),\beta)]\mathrm{d}\beta$ with an error of $O_p(\sqrt{\bar{N}/t_{\text{tune}}} + \bar{N}^{-1})$ for sufficiently well-behaved

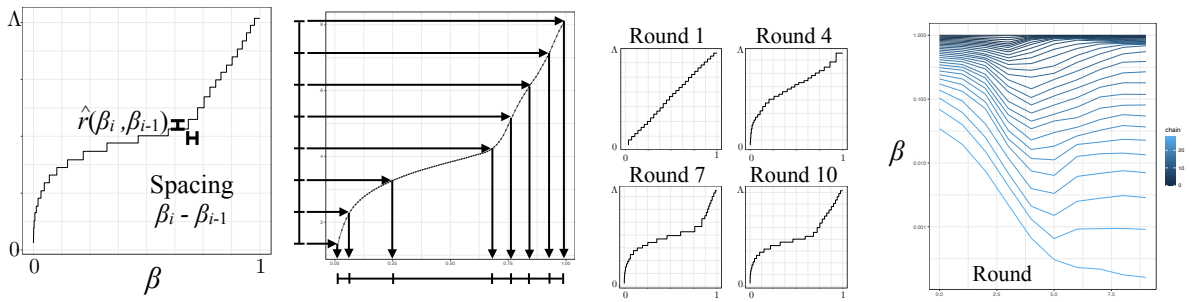

Figure 3.4: Proposed annealing schedule optimization method: example on the Bayesian mixture model of Section 3.4. (1) The cumulative barrier $\Lambda(\cdot)$ is estimated using (3.19) at each point β_k of an initial partition. (2) The cumulative barrier is interpolated using monotonic cubic interpolation, and a new schedule (shown beside the abscissa axis) is obtained by computing the inverse under $\Lambda(\cdot)$ of a regular grid (shown beside the ordinate). (3) The process 1–2 is repeated iteratively. (4) The sequence of annealing schedules obtained as a function of the round (colours represent different grid points in the schedule).

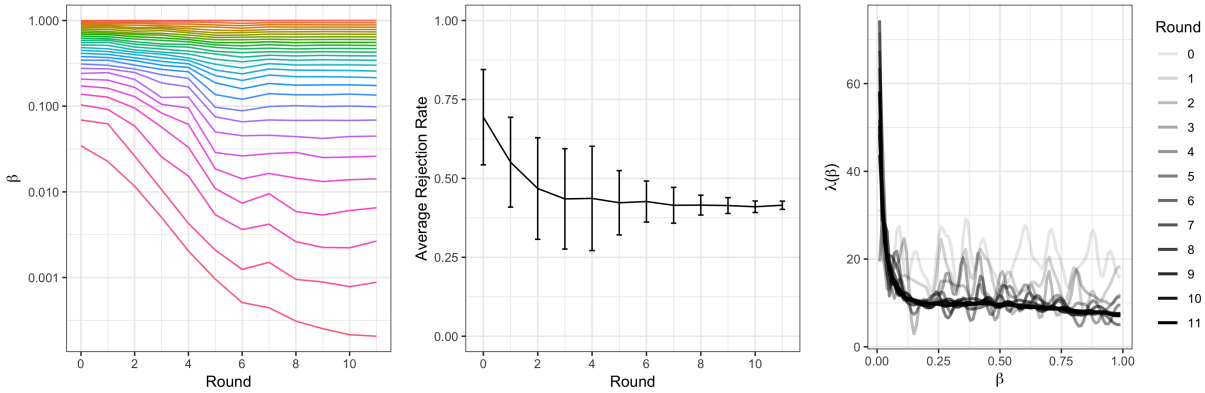

Figure 3.5: A demonstration of the tuning phase in Algorithm 5 ran on a hierarchical Bayesian model applied to the historical failure rates of 5 667 launches for 367 types of rockets (a 369 dimensional problem, see Section 3.4.3 for details). We use $\bar{N} = 30$ cores and 11 schedule optimization rounds, the last one consisting of $t_{\text{tune}} = 5\,000$ and $\hat{\Lambda} = 12.03$. (Left) Progression of the annealing schedule (colours index parallel chains, y-axis, the values β_n for each schedule optimization round, in log scale). (Centre) Progression of the sample mean and standard deviation of empirical rejection probabilities $\{\hat{r}^{(n-1,n)}\}_{n=1}^{N}$. The mean stabilizes quickly, and as the schedule optimization rounds increase, the rejection probabilities converge to the mean as desired. (Right) Progression of the estimated $\hat{\lambda}(\beta)$ as a function of the schedule optimization round.

Algorithm 5 NRPT

Input: Initial state $\boldsymbol{x}_0$, annealing path π, maximum chains $\bar{N}$, tuning budget t_{tune}, sampling budget t_{sample}

▷ Tuning phase

▷ Initialize annealing schedule of size $\bar{N}+1$ (e.g. uniform)

$\mathcal{B}_{\bar{N}} \leftarrow (0, 1/\bar{N}, \ldots, 1)$

▷ Total number of adaptive rounds

`maxRound` $\leftarrow \log_2(t_{\text{tune}})$

$t \leftarrow 1$

for `round` **in** 1, 2, ..., `maxRound` **do**

▷ Approximate rejection rate using Algorithm 1

$\{r^{(n-1,n)}\} \leftarrow \text{DEO}(\pi, \mathcal{B}_N, t)$

▷ Approximate communication barrier using Algorithm 4

$\lambda, \Lambda \leftarrow \text{CommunicationBarrier}(\{r^{(n-1,n)}\}, \mathcal{B}_{\bar{N}})$

▷ Approximate optimal schedule using Algorithm 3

$\mathcal{B}_{\bar{N}} \leftarrow \text{UpdateSchedule}(\lambda, \bar{N})$

▷ Rounds use an exponentially increasing number of scans

$t \leftarrow 2t$

end for

▷ Sampling phase

▷ Optimal number of chains

$N \leftarrow 2\Lambda$

▷ Optimal schedule

$\mathcal{B}_N \leftarrow \text{UpdateSchedule}(\lambda, N)$

▷ Optimal number of copies

$k \leftarrow \bar{N}/(N+1)$

for $1, \ldots, k$ **do**

$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t_{\text{sample}}}) \leftarrow \text{DEO}(\pi, \mathcal{B}_N, t_{\text{sample}})$

return $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t_{\text{sample}}})$

end for

functions f. Examples of such quantities include log-normalizing constants (Kirkwood, 1935; Gelman and Meng, 1998; Xie et al., 2011), KL-divergence (Dabak and Johnson, 2002), and the Fisher-Rao length (Amari, 2016) between π_0 and π_1.

For example, by taking a Riemann sum in the thermodynamic integration identity (Kirkwood, 1935; Gelman and Meng, 1998), the log-normalizing constant $\log Z(1)$ of π can be approximated using

$$\begin{aligned}\log Z(1) &= \log Z(0) - \int_0^1 \mu(\beta)\mathrm{d}\beta \\ &= \log Z(0) - \sum_{n=1}^{N} \mu(\beta_{n-1})(\beta_n - \beta_{n-1}) + O(N^{-1}),\end{aligned} \tag{3.21}$$

where $Z(0)$ the normalizing constant of π_0 assumed to be known and $\mu(\beta) = \mathbb{E}_\beta[V]$. For each round of Algorithm 5 with schedule $\mathcal{B}_N$ for n scans we substitute a Monte Carlo estimate $\hat{\mu}(\beta_k)$ for $\mu(\beta_k)$ into (3.21) to get a consistent estimator $\log \hat{Z}(1)$ for $\log Z(1)$

$$\log \hat{Z}(1) \approx \log Z(0) - \sum_{n=1}^{N} \hat{\mu}(\beta_{n-1})(\beta_n - \beta_{n-1}),$$

with error $O_p(\sqrt{N/T} + N^{-1})$. Figure 3.6 (bottom) in Section 3.4.1 shows how the estimate for $\log \hat{Z}(1)$ evolves with the number of rounds in Algorithm 5 for 16 different models.

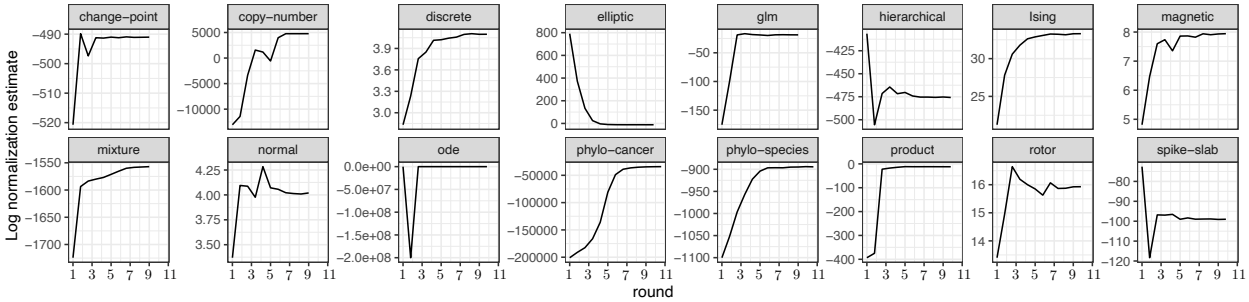

Figure 3.6: Progression of the log normalization constant estimates for 16 different models. Estimates are derived from Section 3.3.5 produced during each NRPT adaptation round.

3.4 Experiments

3.4.1 Empirical behaviour of the schedule optimization method

We applied the proposed NRPT algorithm to 16 models from the statistics and physics literature to demonstrate its versatility. They include 9 Bayesian models ranging from simple standard models such as generalized linear models and Bayesian mixtures, to complex ones such as cancer copy-number calling, ODE parameter estimation, spike-and-slab classification and two types of phylogenetic models. This is complemented by three models from statistical mechanics and four artificial models. The datasets considered include eight real datasets spanning diverse data-types and size, e.g. state-of-the-art measurements such as whole-genome single-cell sequencing data (494 individual cells from two types of cancer, triple negative breast cancer and high-grade serous ovarian (Dorri et al., 2020)) and mRNA transfection time series (Leonhardt et al., 2014), as well as more

conventional ones such as mtDNA data and various feature selection/classification datasets. See Table 3.1 and references therein for details. All models and algorithms are implemented in the Blang modelling language (Bouchard-Côté et al., 2021).

In each of the 16 models and for each round of Algorithm 5 (line 4), we computed the mean swap acceptance probability across all neighbour chains. We then summarized these means across chains using a box plot. The equi-acceptance objective function of Section 3.3.1 can be visually understood as collapsing this box plot into a single point. We show in Figure 3.7 (top) the progression of these swap acceptance probabilities. In all examples considered, equi-acceptance is well approximated within 10 rounds.

Within the range of models considered, we observed a diversity of local barriers λ estimated by Algorithm 5 (Figure 3.7 (bottom)). Most statistical models exhibit a high but narrow peak in the neighbourhood of the reference ($\beta = 0$). However, a subset of models including statistical models (mixture, ode, phylo-cancer, spike-slab) and physics models (Ising, magnetic, rotor) exhibit additional peaks away from $\beta = 0$. See Figure 2.1 for the corresponding global barriers Λ estimates as a function of the rounds, and the resulting schedule generator.

Reproducibility.

To make our NRPT 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 local exploration kernels, namely slice sampling with exponential doubling followed by shrinking (Neal, 2003). Scripts documenting replication of our experiments are available at `https://github.com/UBC-Stat-ML/ptbenchmark`.

Multi-core implementation.

We use lightweight threads (Friesen, 2015) to parallelize both the local exploration and communication phases, as shown in Algorithm 1. We use the algorithm of Leiserson et al. (2012) as implemented in Steele and Lea (2013) to allow each PT chain to have its own random stream. This technique

Model (and dataset when applicable)	n	d	$\hat{\Lambda}$	N
Change point detection (text message data, Davidson-Pilon (2015))	74	3	4.0	20
Copy number inference (whole genome ovarian cancer data, Section 3.4.5)	6 206	30	13.0	50
Discrete multimodal distribution (Section 3.2.6)	N/A	3	0.4	30
Weakly identifiable *elliptic* curve (Section 3.4.3)	N/A	2	4.4	30
General Linear Model (*GLM*) (Challenger O-ring dataset, Dalal et al. (1989))	23	2	3.3	15
Bayesian *hierarchical* model (historical rocket failure data, Section 3.4.3)	5 667	369	12.0	30
Ising model (Section 3.2.6)	N/A	25	3.1	30
Ising model with *magnetic* field (Section 3.2.6)	N/A	25	2.3	30
Bayesian *mixture* model (Section 3.4.4)	300	305	8.2	30
Bayesian *mixture* model (subset of 150 datapoints)	150	155	5.5	20
Isotropic *normal* distribution	N/A	5	1.4	30
ODE parameters (mRNA data, Leonhardt et al. (2014))	52	5	6.4	50
Phylogenetic inference (single cell breast cancer data, Dorri et al. (2020))	192 763	192 765	88	300
Phylogenetic species tree inference (mtDNA, Hayasaka et al. (1988))	249	10 395	7.1	30
Unidentifiable *product* parameterization (Section 3.4.3)	100 000	2	3.7	15
Rotor (XY) model, Hsieh et al. (2013)	N/A	25	3.3	40
Spike-and-slab classification (RMS Titanic passengers data (Hind, 2019))	200	19	4.7	30

Table 3.1: Summary of models used in the experiments, with the number of observations n (when applicable), the number of latent random variables d, estimated $\hat{\Lambda}$, and the default number of chains used. An abbreviation for each is shown in italic. The model-specific command line options used for all experiments is available at `https://github.com/UBC-Stat-ML/blangDemos/blob/master/src/main/resources/demos/models.csv`. The same file also documents the location of the probabilistic programming source code and precise command line arguments for the 16 models.

avoids any blocking across threads and hence makes the inner loop of our algorithm embarrassingly parallel in N. Moreover, the method of Leiserson et al. (2012) 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 number 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 3.4.3 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).

Selection of π_0

In our experiments all Bayesian models considered have proper priors, and hence we set π_0 to the prior in all these situations. The local exploration kernel in this case consists in an independent draw from the prior (performed by sorting the latent random variables according to a linearization of the partial order induced by the directed graphical model, and sampling their values according to these sorted laws).

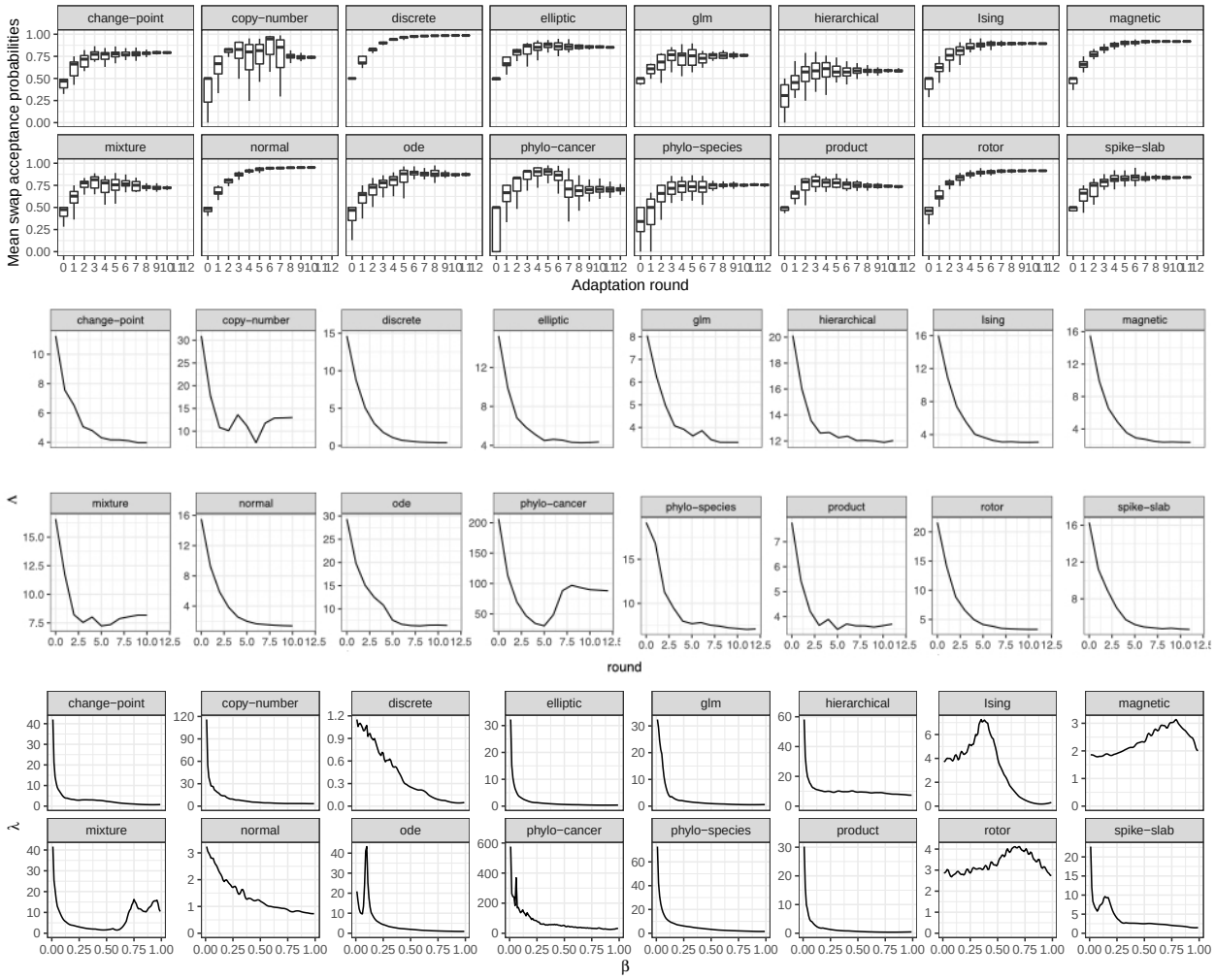

Figure 3.7: Empirical behaviour of NRPT on 16 models. Top: distribution of acceptance rates observed at each round of Algorithm 5. Middle: estimates of the global communication barrier Λ. The abscissa denotes the schedule optimization round. Bottom: estimates of the local communication barrier $\hat{\lambda}$ in the final round.

For the Ising model, we let π_0 denote a product of independent and identically distributed Bernoulli(1/2) random variables, one for each node in the Ising grid. It is then straightforward to interpolate between this product distribution and the Ising model of interest using a geometric average. Independent sampling from π_0 is used for the local exploration kernel at $\beta = 0$.

Similarly, for the rotor (XY) model, we take π_0 to be a product of independent and identically distributed Uniform$(-\pi, \pi)$ random variables, and we use a geometric interpolation between π_0 and the target distribution.

Multimodality of the examples considered

Figures 3.1 (bottom left), 3.11 and 3.12 support the multimodality of two of the examples considered in Section 3.4.3, namely the Ising and Spike-and-Slab examples. Multimodality of other examples considered in Section 3.4.1, 3.4.2, 3.4.4 and 3.4.5 is demonstrated in Figures 3.1, 3.18, 3.15, 3.17, and 3.18. Moreover, Figures 3.1, 3.15, 3.17, and 3.18 demonstrate that using standard MCMC is insufficient to explore these multimodal distributions.

3.4.2 Robustness to ELE violation

To investigate empirically whether the NRPT methodology is robust to the violation of the ELE assumption, we ran NRPT with a range of values for t_{expl} on the models shown in Figure 3.8. Let d_{var} denote the number of variables in each model. We run experiments with $t_{\text{expl}} = 0$, $(1/2)d_{\text{var}}$, d_{var}, $2d_{\text{var}}$, $4d_{\text{var}}$, $\ldots$, $32d_{\text{var}}$ (the only exception is the reference chain ($\beta = 0$), where we always use $t_{\text{expl}} = 1$ since we can get exact samples from π_0). The key quantity used by the NRPT algorithm for schedule optimization is the communication barrier λ. The results shown in Figure 3.8 demonstrate that in all models considered the function λ is reliably estimated even when ELE is severely violated, provided $t_{\text{expl}} > 0$. Similarly, since the estimates of Λ and $\bar{\tau}$ are derived from λ, these quantities can be accurately estimated even when ELE is severely violated. The results shown in Figure 2.4 support that increasing t_{expl} reduces the difference between the theoretical and observed round trip rate. Moreover, Figure 2.4 also demonstrate a second strategy to alleviate ELE violation, which is to increase N while fixing t_{expl}.

In the next experiment shown in Figure 3.9, we compare the two mechanisms available for alleviating ELE violation, namely increasing N and increasing t_{expl}. First consider the black line in Figure 3.9 showing the regime where ELE is best approximated in these experiments. This is in close agreement with the theoretical guidelines developed in Section 3.3.3. At the same time, the light blue lines in Figure 3.9 show that in situations where the local exploration kernel is not easy to parallelize, it can be advantageous to increase t_{expl} and to compensate with a number of chains N higher than 2Λ, leading to an average swap rejection rate $r < 1/2$.

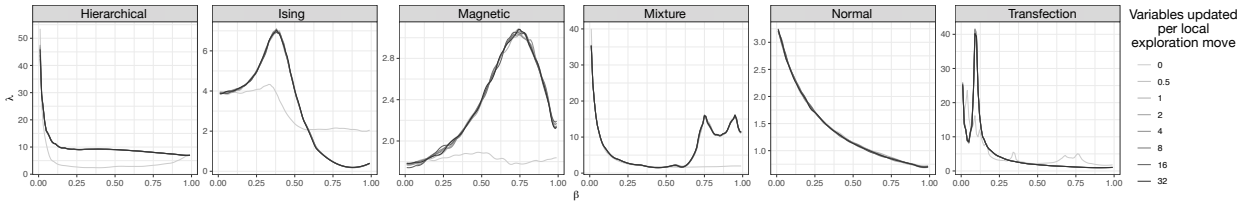

Figure 3.8: Estimate $\hat{\lambda}$ of the local communication barrier λ for different values of $t_{\text{expl}} \geq 0$ and different models (refer to Table 3.1 for model descriptions). The estimates are consistent as long as $t_{\text{expl}} > 0$, even when assumption (A2) is severely violated. Refer to Section 3.4.2 for experiment details.

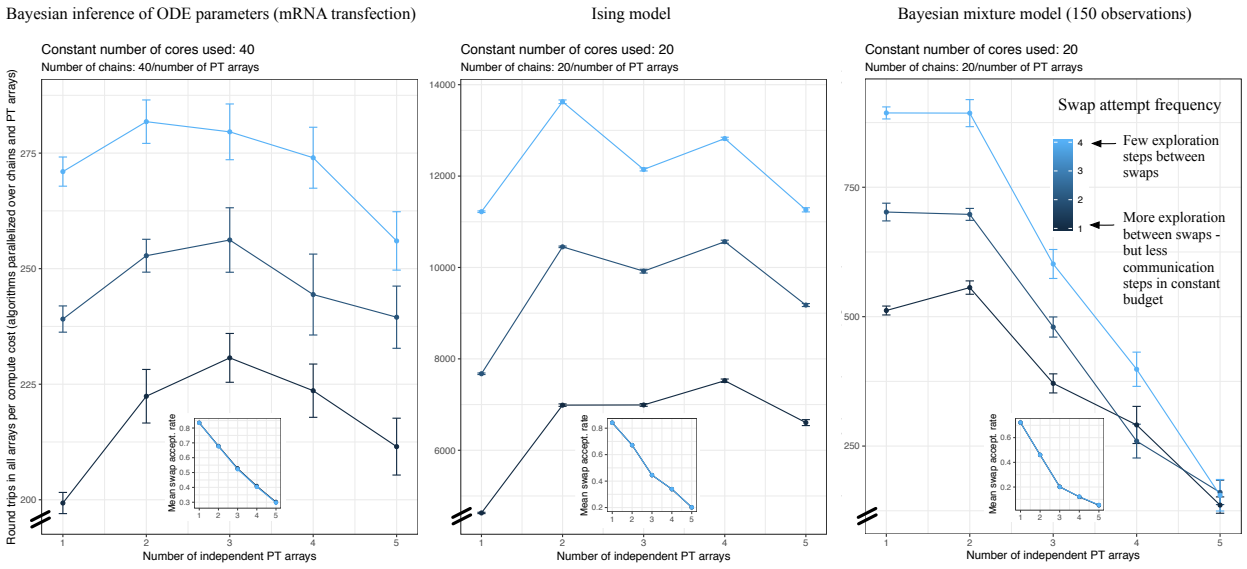

Figure 3.9: Trade-off between number of chains N, number of independent PT algorithms, k, and the frequency at which swaps are attempted ($\propto 1/t_{\text{expl}}$). Each model uses a constant number of cores and set $N + 1 = \bar{N}/k$ chains per independent PT algorithm. Insets: swap acceptance probability for each k. The results in the black lines (where ELE is best approximated) agree with the theory of Section 3.3.3. For the ODE model, $\Lambda \approx 6.4$, $N^* \approx 12.8$ in agreement with $k^* = 3$ observed. For the Ising model, $\Lambda \approx 3.1$, $N^* \approx 6.2$ in agreement with $k^* = 4$ observed. For the Bayesian mixture model, $\Lambda \approx 5.5$, $N^* \approx 11$, in agreement with $k^* = 2$ observed. These results also show that in the context of local exploration moves that are not easily parallelized, it is better to use higher swap attempt frequencies (light blue lines) which achieve their optima at $N > 2\Lambda$ and average swap rejection rate $r < 1/2$.

3.4.3 Comparison with other parallel tempering schemes

We benchmarked the empirical running time of Algorithm 5 compared to previous adaptive PT methods (Atchadé et al., 2011; Miasojedow et al., 2013). The methods we considered are: (1) the stochastic optimization adaptive method for reversible schemes proposed in Atchadé et al. (2011);

(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 Miasojedow et al. (2013); (3) our non reversible schedule optimization scheme (NRPT); and finally, (4) our scheme, combined with a better initialization based on a preliminary execution of a sequential Monte Carlo algorithm (more precisely, based on a "sequential change of measure," labelled SCM, as described in Del Moral et al. (2006)), 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). We benchmarked the methods on four models: (a) a 369-dimensional hierarchical model applied to a dataset of rocket launch failure/success indicator variables (McDowell, 2019); (b) a 19-dimensional Spike-and-Slab variable selection model applied to the RMS Titanic Passenger Manifest dataset (Hind, 2019); (c) A 25-dimensional Ising model from Section 3.2.6 ($M = 5$); (d) a 9-dimensional model for an end-point conditioned Wright-Fisher stochastic differential equation (see, e.g., Tataru et al. (2017)).

All baseline methods are implemented in Blang (`https://github.com/UBC-Stat-ML/blangSDK`), the same probabilistic programming language 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 Atchadé et al. (2011) and Miasojedow et al. (2013) are based on reversible PT together with two different flavours of stochastic optimization to adaptively select the annealing schedule. In Atchadé et al. (2011), the chains are added one by one, each chain targeting a swap acceptance rate of 23% from the previous one. In Miasojedow et al. (2013), this scheme is modified in two ways: first, all annealing parameters are optimized simultaneously, and second, a different update for performing the stochastic optimization is proposed. 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 Miasojedow et al. (2013). Specifically, both Atchadé

et al. (2011) and Miasojedow et al. (2013) 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 of Atchadé et al. (2011) uses the update $\beta'_n = \beta(1 + \exp(\rho_n))^{-1}$, whereas the work of Miasojedow et al. (2013) 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 Atchadé et al. (2011) use $\gamma_n = (n+1)^{-1}$, Miasojedow et al. (2013) suggest to use $\gamma_n = (n+1)^{-0.6}$. The experiments confirm that the latter algorithm is more stable.

Description of models and datasets

We benchmark the methods described above on the following four models. First, a hierarchical model applied to a dataset of rocket launch failure/success indicator variables (McDowell, 2019). We organized the data by types of launcher, obtaining $5,667$ launches for 367 types of rockets (processed data available at `https://github.com/UBC-Stat-ML/blangDemos/blob/master/data/failure_counts.csv`). Each type is associated with a Beta-distributed parameter with parameters tied across rocket types, with the likelihood given by a Binomial distribution (full model specification available at `https://github.com/UBC-Stat-ML/blangDemos/blob/master/src/main/java/hier/HierarchicalRockets.bl`). The second model is a Spike-and-Slab variable selection model applied to the RMS Titanic Passenger Manifest dataset (Hind, 2019). The preprocessed data is available at `https://github.com/UBC-Stat-ML/blangDemos/tree/master/data/titanic`. The data consist in binary classification indicators for the survival of each individual passenger as well as covariates such as age, fare paid, etc. We used a Spike-and-Slab prior with a point mass at zero and a Student-t continuous component (full model specification available at `https://github.com/UBC-Stat-ML/blangDemos/blob/master/src/main/java/glms/SpikeSlabClassification.bl`). Third, we used the Ising model from Section 3.2.6. Finally, we also used an end-point conditioned Wright-Fisher stochastic differential equation (see, e.g., Tataru et al. (2017)). For this last model we used synthetic data generated by the model. The specification of this last model is available at `https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/validation/internals/fixtures/Diffusion.bl`.

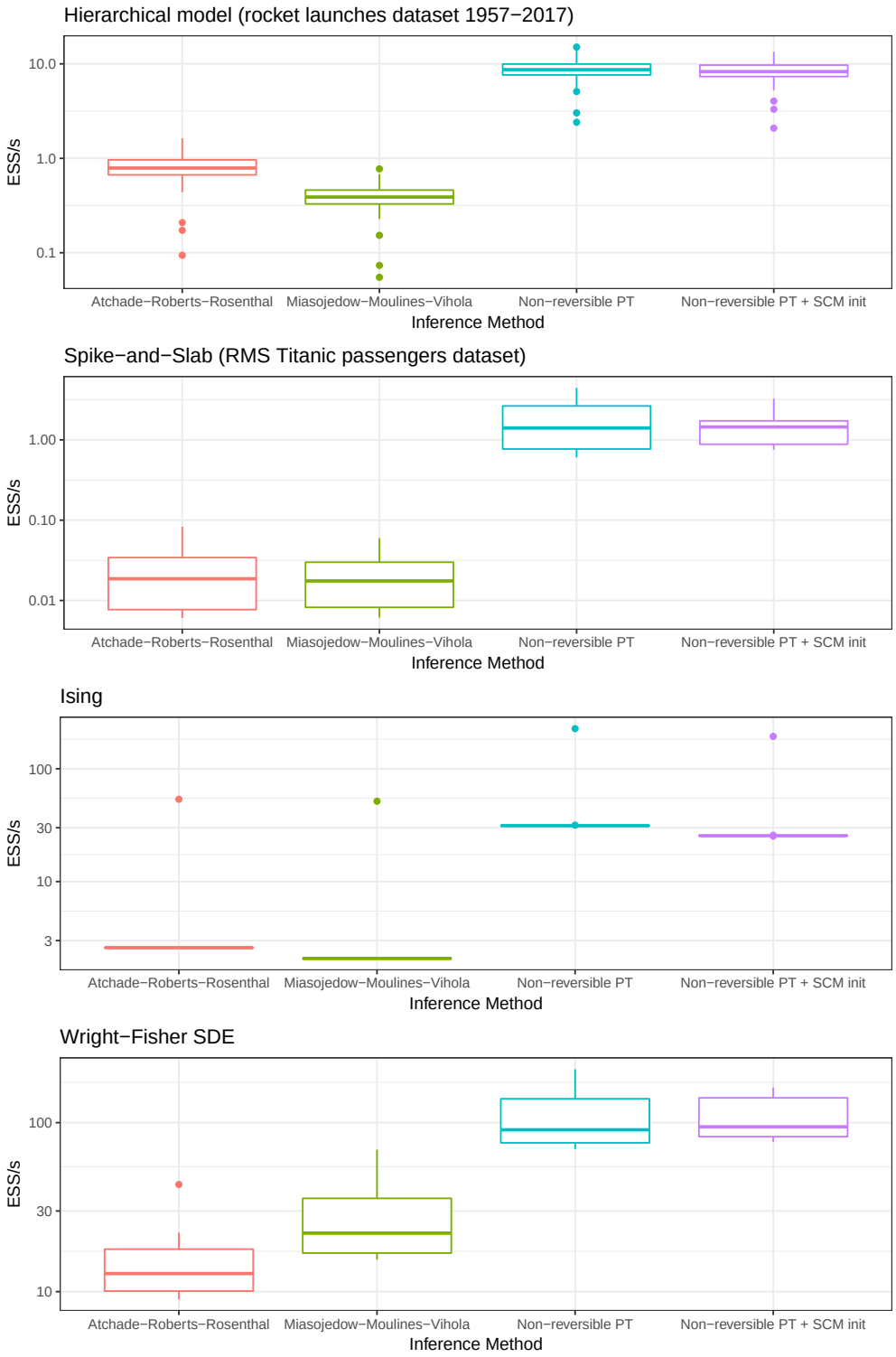

Figure 3.10: Effective Sample Size (ESS) per second (ordinate, in log scale) for four PT methods (abscissa). The four facets show results for the four models described in Section 3.4.3.

Model (and dataset when applicable)	Method	N	$\hat{s}$ (%)	$\hat{\Lambda}$
Spike-and-slab classification	NRPT+SCM	15	67.6	4.54
(RMS Titanic passengers data, Hind (2019))	ARR	6	36.3	3.82
	MMV	6	39.9	3.61
	NRPT	15	67.2	4.59
Bayesian hierarchical model	NRPT+SCM	34	63.7	11.96
(historical rocket failure data, McDowell (2019))	ARR*	2*	0*	1*
	MMV	15	33.6	9.96
	NRPT	34	63.4	12.08
Wright-Fisher diffusion	NRPT+SCM	15	75.2	3.47
	ARR	5	43.8	2.81
	MMV	5	43.7	2.82
	NRPT	15	75.5	3.42
Ising model	NRPT+SCM	15	78.2	3.05
	ARR	4	37.9	2.48
	MMV	4	36.9	2.52
	NRPT	15	78.5	3.01

Table 3.2: Summary statistics for the experiments in Section 3.4.3. The row marked with a star indicates failed optimization of one of the stochastic optimization schemes on the Bayesian hierarchical model.

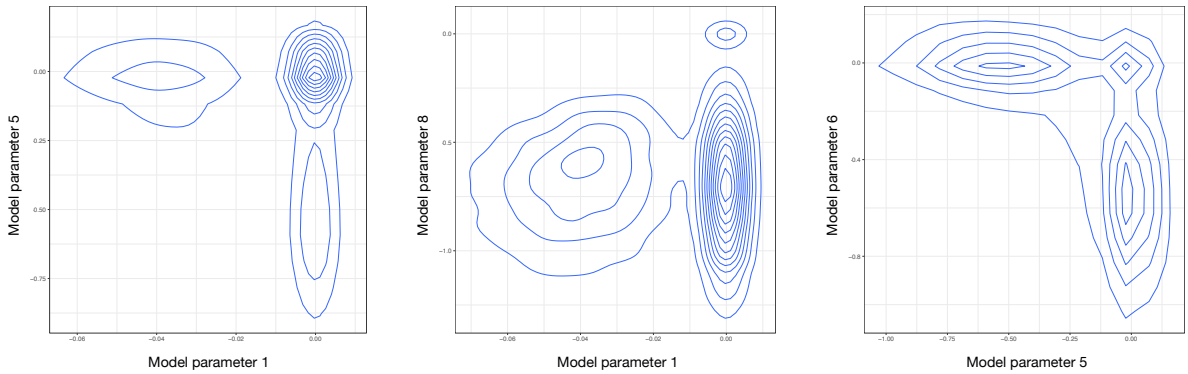

Figure 3.11: Examples of multimodality encountered in the Spike-and-Slab model applied to the RMS Titanic passenger dataset. Joint posterior distribution of the following regression parameters (1) `passengerAge` and `passengerClass`; (2) `passengerAge` and `SiblingsSpousesAboard`; (3) `passengerClass` and `passengerClass2`, the latter corresponding to an artificially duplicated regressor used to investigate the effect of co-linearity on the posterior approximation.

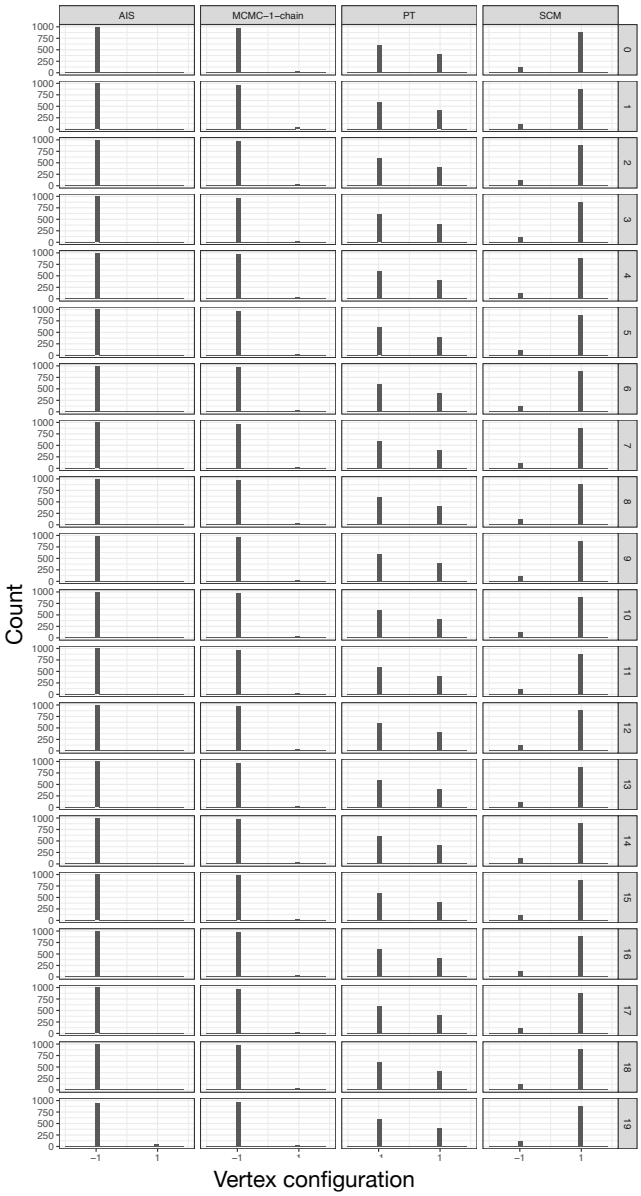

Figure 3.12: Example of multimodality in an Ising model. Each facet is a posterior probability mass function, facet rows index the first 20 vertices of the Ising graph, facet columns, four posterior approximation methods. By symmetry, each marginal should place equal mass at spins -1 and 1. The multimodality is only correctly captured by our proposed algorithm (NRPT, third column, wall clock time of 7.250s, $N = 10$). For the other methods, the posterior approximation either misses the other mode completely, namely when using AIS (left column, wall clock time of 5.723s, 1000 particles, relative ESS adaptive annealing schedule) or single-chain MCMC (second column, wall clock time of 1.439s), or, for annealed SMC (fourth column, wall clock time of 8.269s, 1000 particles, relative ESS adaptive annealing schedule), detects the multimodality but not the modes' respective proportions. All wall clock times are reported for a 2.8 GHz Intel Core i7.

3.4.4 Mixture models

Bayesian analysis of mixture models can give rise to a label-switching symmetry, leading to a multimodal posterior distribution. We consider a Bayesian mixture model with two mixture components. The likelihood for each component is a normal distribution with a non-conjugate Uniform$(0, 100)$ prior on the standard deviation and a normal prior on the means (standard deviation of 100). We placed a uniform prior on the mixture proportion. We used simulated data generated from the model. While the mixture membership indicator latent random variables can be marginalized in this model, we sample them to make the posterior inference problem more challenging. Sampling these mixture membership random variables is representative of more complex models from the Bayesian non-parametric literature where marginalization of the latent variable is intractable; for example, this is the case for the stick-breaking representation of general completely random measures (Zhu et al., 2020).

In addition to the three MCMC methods described in the main text, we also ran baselines based on SMC and AIS, which are popular methods to explore complex posterior distributions. These methods also depend on the construction of a sequence of annealed distributions from prior to posterior. For the SMC and AIS baselines, to select the sequence of distributions we used an adaptive scheme based on relative ESS as described in Zhou et al. (2016). Diagnostics of the adaptation are shown in Figure 3.16. These methods were parallelized at the particle level. We set the number of particles to achieve a similar running time compared to NRTP, namely 2 000 particles for SMC and 2 500 particles for AIS. We found that the quality of the posterior approximation was highly dependent on performing several rounds of rejuvenations on the final particle population. The wall-clock time with 5 and 20 rounds of rejuvenation (SMC-5, SMC-20) was comparable to NRPT (1.778min, 2.302min) however the posterior approximation is markedly poor compared to NRPT. With 100 rounds of rejuvenation, the posterior approximation matches closely that of NRPT, however this brings the computation cost to 5.064min. AIS did not perform well since the weights were highly unbalanced in the last iteration, effectively resulting in an approximation putting all mass to a single particle.

In Figure 3.13 we compare the following inference methods on a label-switching posterior distribution: our proposed algorithm (NRPT), an MCMC run based on a single chain (i.e. the

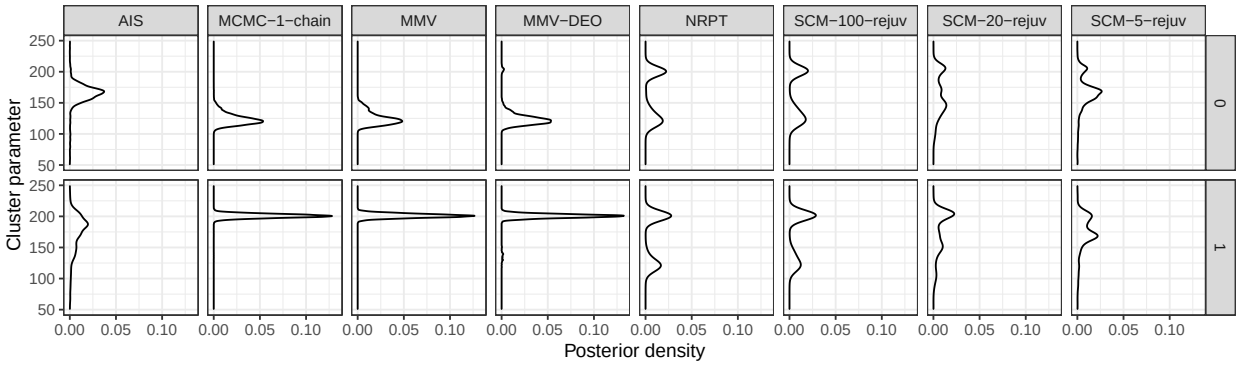

Figure 3.13: Mixture model example: eight approximations of the posterior distributions, showing for each approximation two of the 155 latent random variables, namely the two component mixtures' mean parameters. Since the two latent random variables are exchangeable a posteriori (by label switching), the true marginal posterior distributions are identical for the two random variables. Amongst the approximation methods, multimodality is only captured by our algorithm (NRPT, 1.508min) and by a state-of-the-art sequential Monte Carlo combined with several rounds of rejuvenation (SCM-100, 5.064min). The benchmark is conservative in that all competing methods use a computational budget (in terms of both parallelism and wall clock time) greater or equal than NRPT.

exploration kernel alone), the stochastic optimization method of Miasojedow-Moulines-Vihola (MMV), DEO but optimized using MMV (MMV-DEO), Annealed Importance Sampling (AIS) (Neal, 2001), and a sequential Monte Carlo based on a sequence of annealed distributions (Del Moral et al., 2006) (labelled SCM as before). For both SCM and AIS, we use the adaptive scheme of Zhou et al. (2016) (see Figure 3.14 for diagnostics of the SCM adaptation). The number of iterations are set so that the method with the smallest wall clock time is our proposed algorithm (NRPT: 1.508min, MMV: 2.019min, MMV-DEO:1.665min, AIS: 1.887min, SCM: 1.778min–5.064min). For all methods, timing includes the time spent to perform schedule optimization.

Amongst MCMC methods, only NRPT correctly captures the multimodality of the target distribution. This is confirmed by the trace plots of the three MCMC methods, shown in Figure 3.15. MMV automatically selected $N = 8$ by targeting a swap acceptance probability of 23%. Post-adaptation, the swap acceptance probability was 27%. For NRPT we use $N = 30$ but to avoid penalizing MMV for a lack of parallelism, we limited all methods to use no more than 8 threads. After schedule optimization, NRPT estimates the global communication barrier to a value of $\hat{\Lambda} \approx 8$, and the average swap acceptance probability was 72% (see Figure 3.16 for more NRPT diagnostics).

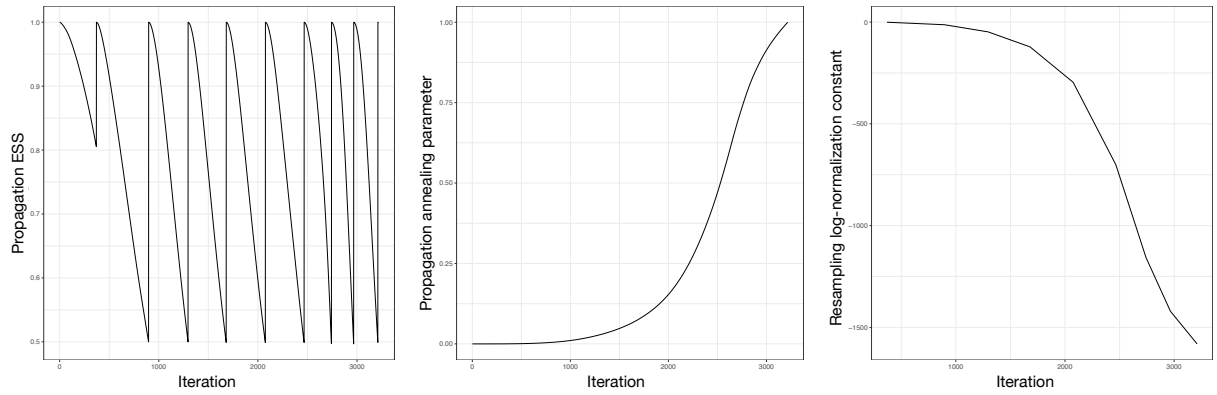

Figure 3.14: Diagnostics for the adaptive annealed SMC method (SCM), used as a benchmark on the Bayesian mixture problem. From left to right: (1) ESS as a function of the SMC iteration. Resampling is performed when the ESS drops below 1/2; (2) annealing parameter as a function of the SMC iteration; (3) log normalization estimates at each resampling step.

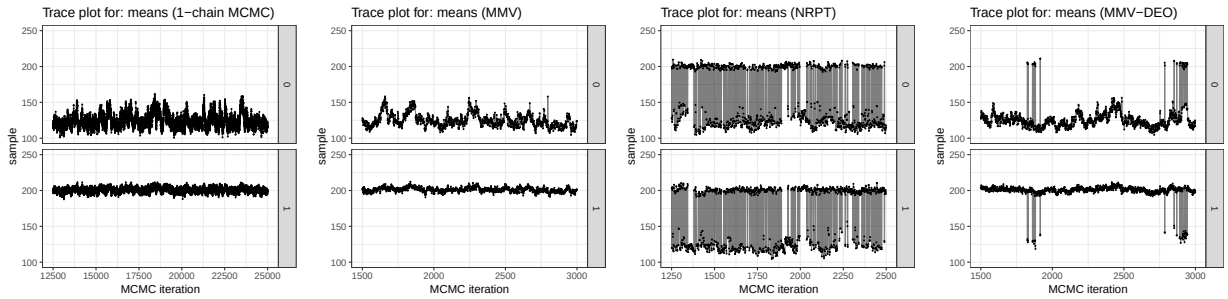

Figure 3.15: Mixture modelling example: post burn-in trace plots of two model parameters (facet rows) for four MCMC methods considered (facet columns). The two parameters correspond to the location parameters of two exchangeable clusters. Correct MCMC exploration of this unidentifiable model requires label switching, providing a test bed for MCMC over multimodal targets. All methods use a computational budget lower or equal to NRPT's.

3.4.5 Multimodality arising from single cell, whole genome copy number inference

In this section we describe an application to copy number inference in which multimodality arises from the unknown ploidy of a cancer cell. The likelihood of this model is based on a hidden Markov model over $n = 6\,206$ observations, and after analytic marginalization of the corresponding 6 206 hidden states, the multimodal sampling problem is defined over $d = 30$ remaining latent variables. The model is described in (Syed et al., 2021a, Appendix I.4.5), in this section we summarize the results concerning the performance of the algorithms.

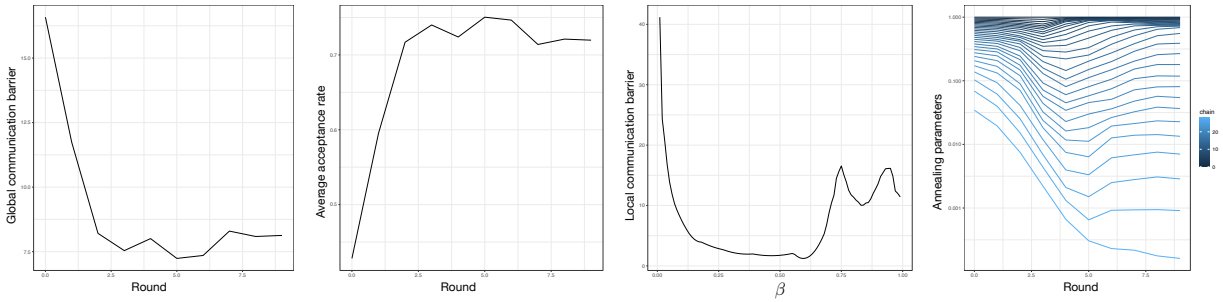

Figure 3.16: Diagnostics for NRPT (Algorithm 5) on the mixture modelling example. From left to right: (1) Estimate of the global communication barrier $\hat{\Lambda}$ as a function of the schedule optimization round. (2) Average swap acceptance probability across the 30 chains as a function of the schedule optimization round. The final value is 72%. (3) Estimated communication barrier λ output by NRPT. The two peaks can be interpreted as transition points where the cluster membership indicator variables go from disorganized (all cluster membership variables are i.i.d. at $\beta = 0$), to all taking the same value within a cluster. The two clusters having different number of data points and parameters induce two distinct transition points. The empirical behaviour observed is analogous to the phase transition found in statistical mechanics models such as the Ising model. Investigation of possible phase transitions in clustering models would be an interesting future direction of investigation, although somewhat orthogonal to this work. (4) Learning curves for the annealing parameters (ordinate axis, log scale) for the 30 NRPT chains (colours) as a function of the schedule optimization round (abscissa).

We compare the quality of posterior approximations from four samplers on the High-Grade Serous Ovarian cancer dataset from Dorri et al. (2020). The four methods compared are: the stochastic optimization adaptive PT method MMV (adaptation selected 7 chains, average post-adaptation swap acceptance, 28.7%), DEO but optimized using MMV (MMV-DEO, adaptation selected 11 chains, average post-adaptation swap acceptance, 26.9%) our proposed algorithm algorithm (NRPT with 25 and 50 chains, average post-optimization swap acceptance of respectively 48.8% and 73.5%), and inference based on a single MCMC chain. As in Section 3.4.4, to avoid overly penalizing MMV for a lack of parallelism, we limited all methods to use no more than 8 threads, hence obtaining the following wall clock running times including schedule optimization when applicable: MMV, 15.91h; MMV+DEO, 14.33h; 25 chains NRPT, 8.809h; 50 chains NRPT, 15.35h; 1 chain MCMC, 2.91h. The running time of MMV was dominated by adaptation (88% of the wall clock time), and is higher than the schedule optimization time used by NRPT (67%).

Refer to Figure 3.17 for a comparison of the trace plots between all three methods. Since the experimental protocol provides only proportionality between read counts and copy number, not

absolute expected read count for a given copy number, the likelihood does not distinguish between a given copy number profile, and another one obtained by genome duplication of the same profile (refer to (Syed et al., 2021a, Appendix I.4.5) for details). The prior favours the former, but when local copy number events occur after the genome duplication it may be possible to detect recently evolved genome duplication. Challenging multimodality arises when these subsequent events involve only small regions. In such case, the tension between the prior distribution favouring low ploidy and the noisy observation favouring higher ploidy creates a multimodal posterior distribution. We use data from one such cell in the following experiments to illustrate a realistic multimodal inference problem.

The multimodality of the posterior distribution was successfully captured by NRPT but not by a single-chain MCMC nor by the stochastic optimization method MMV (Figure 3.17). MMV failed to achieve any round trip because the learnt schedule is highly suboptimal: while the mean swap acceptance probability across chains is 28.7%, the minimum across chains is nearly zero due to the noise in the optimization procedure. The difficulty of exploring this particular multimodal target is compounded by the fact that the state space of each chromosome's copy number is random, being upper-bounded by random variables m_c corresponding to each chromosome c. So in order to perform a jump doubling the cell's ploidy the sampler has to increase a large number of discrete variables m_c simultaneously. This is reflected in the posterior distribution of the model variables denoted m_c, as seen in trace plots in Figure 3.17. Hence the local exploration kernel, which in this example samples the variables m_c one at the time, is insufficient to jump mode, and only excursions through the prior can achieve mode jumping, via a regeneration based on sampling from the prior at $\beta = 0$.

NRPT used 50 chains, estimated $\hat{\Lambda} = 13$ (see Figure 3.19), and converged quickly in this large scale example, performing drastic changes in the annealing schedule in the first 8 rounds then relatively little changes in rounds 8 through the final tenth round (Figure 3.19).

Comparing the performance of the two NRPT variants, we estimated a round trip rate of 9×10^{-4} for 25-chain, and 6×10^{-3} for 50-chain. Notice the increase being more than a factor two supports the use of large N for exploration of complex multimodal posterior distributions. Neither MMV nor MMV-DEO achieved any round trips in the post-burn-in iterations.

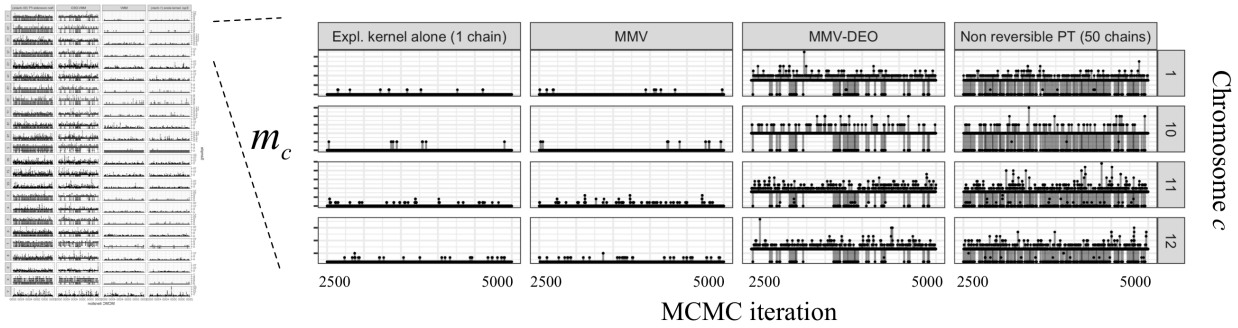

Figure 3.17: Trace plots for the copy number inference problem. Each facet shows a post burn-in trace plot for one of the random variables m_c. Facet rows are indexed by chromosomes. Facet columns are indexed by MCMC approximation methods. The trace plots show that only NRPT frequently jumps between the two modes. MMV and the exploration kernel alone do not achieve any jump, due to the required concerted changes in a large number of discrete variables.

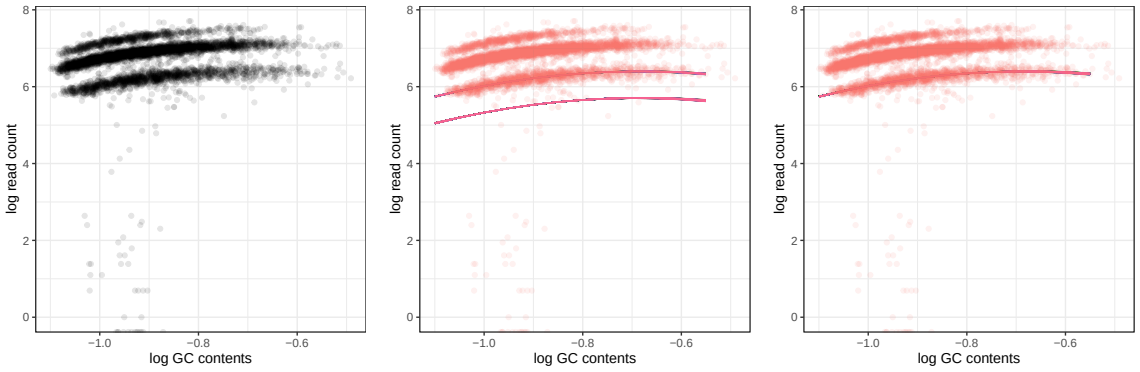

Figure 3.18: Copy number inference: inputs and outputs. From left to right: (1) raw data, where each dot represent a genomic bin i, c, with its abscissa showing $\log g_{i,c}$ and its ordinate, $\log y_{i,c}$. (2) posterior distribution obtained from NRPT for the random function $f(\theta, \cdot)$ obtained from NRPT. One translucent line $f(\theta_n, \cdot)$ is drawn for each sampled θ_n. Notice the bi-modality of the posterior distribution. (3) The approximation of the same posterior distribution based on a single chain misses one of the modes.

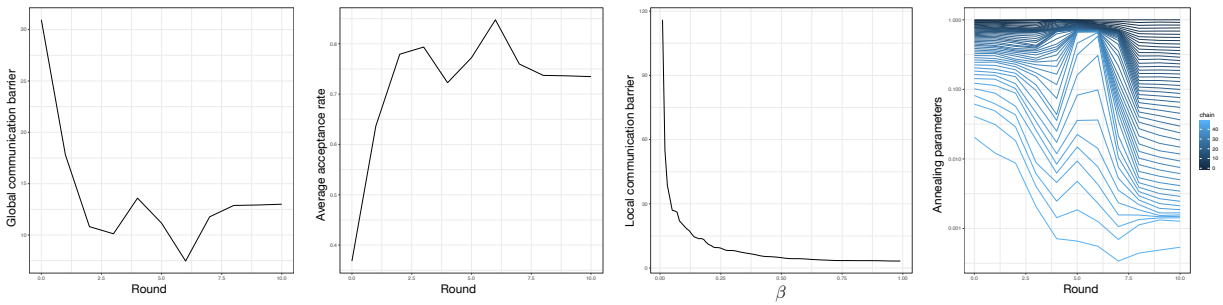

Figure 3.19: NRPT diagnostics on the copy number inference example. From left to right: (1) Estimate of the global communication barrier $\hat{\Lambda}$ as a function of the scehdule optimization round. (2) Average swap acceptance probability across the 50 chains as a function of the schedule optimization round. (3) Estimated communication barrier λ output by NRPT. (4) Learning curve for the annealing parameters (ordinate axis, log scale) for the 50 NRPT chains (colours) as a function of the schedule optimization round (abscissa).

Chapter 4

Parallel tempering on optimized paths

One geometry cannot be more true than another; it can only be more convenient.

— Jules Henri Poincare

4.1 Motivation

The previous chapter showed that non-reversible PT is guaranteed to dominate its classical reversible counterpart. Moreover, adding more chains in the non-reversible regime does not lead to performance collapse. However, even with these more efficient non-reversible PT algorithms, we established that the improvement in round trip rates brought by higher parallelism would asymptote to a fundamental limit $\tau_\infty = (2 + 2\Lambda)^{-1}$ controlled by the global communication barrier Λ. The communication barrier Λ measures the difficulty of communication between π_0 and π_1 and represents a limitation of non-reversible PT that cannot be improved upon by increasing the number of chains or tuning the schedule. Therefore when Λ is large, NRPT with a well-tuned schedule with many chains will still suffer poor performance. For example, this can happen when π_0 and π_1 are nearly mutually singular. A typical case is where the target is a Bayesian posterior distribution, the reference is the prior—for which i.i.d. sampling is typically possible—and the prior is misspecified.

We can naturally ask if it is possible to improve upon the optimal round trip rate τ_∞, theoretically and empirically established in Chapter 3. In this chapter, we show that the answer to this question is surprisingly positive. We will show that by generalizing the class of paths interpolating between π_0 and π_1 from linear to nonlinear, the global communication barrier can be broken, leading to substantial performance improvements. Notably, the nonlinear path used to demonstrate this breakage is computed using a practical algorithm that can be used in any situation where PT is applicable.

The following proposition demonstrates that the traditional linear path $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^\beta$ suffers

from an arbitrarily suboptimal global communication barrier even in simple examples with Gaussian reference and target distributions. Therefore, upon decreasing the variance of the reference and target while holding their means fixed, the traditional linear annealing path obtains an exponentially smaller asymptotic round trip rate than the optimal path of Gaussian distributions. Figure 4.1 provides an intuitive explanation. The standard path (top) corresponds to a set of Gaussian distributions with mean interpolated between the reference and target. Reducing the variance of the reference and target also reduces the variance of the distributions along the path. For any fixed N, these distributions become nearly mutually singular, leading to arbitrarily low round trip rates. The solution to this issue (bottom) is to allow the distributions along the path to have increased variances, thereby maintaining mutual overlap and the ability to swap components with a reasonable probability. This motivates the need to design more general annealing paths. In the following, we introduce the precise general definition of an annealing path, an analysis of path communication efficiency in parallel tempering, and a rigorous formulation of—and solution to—the problem of tuning path parameters to maximize the round trip rate.

This chapter will develop a theoretical analysis of parallel tempering algorithms based on general nonlinear paths and their geometry. We will use this to understand the properties of the communication barrier and how it changes and develop a practical algorithm to tune the path to improve the performance of PT. To see an example of a path optimized using our algorithm see Figure 4.1 (bottom).

Proposition 9. *Suppose the reference and target distributions are* $\pi_0 = \mathcal{N}(\mu_0, \sigma^2)$ *and* $\pi_1 = \mathcal{N}(\mu_1, \sigma^2)$*, and define* $z = |\mu_1 - \mu_0|/\sigma$*. Then as* $z \to \infty$*,*

1. *the path* $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^\beta$ *has* $\tau_\infty = \Theta(1/z)$*, and*
2. *there exists a path of Gaussian distributions with* $\tau_\infty = \Omega(1/\log z)$*.*

We will prove a more general result in Section 4.5.9.

4.1.1 Literature review

Beyond parallel tempering, several methods to approximate intractable integrals rely on a path of distributions from a reference to a target distribution, and there is a rich literature on the

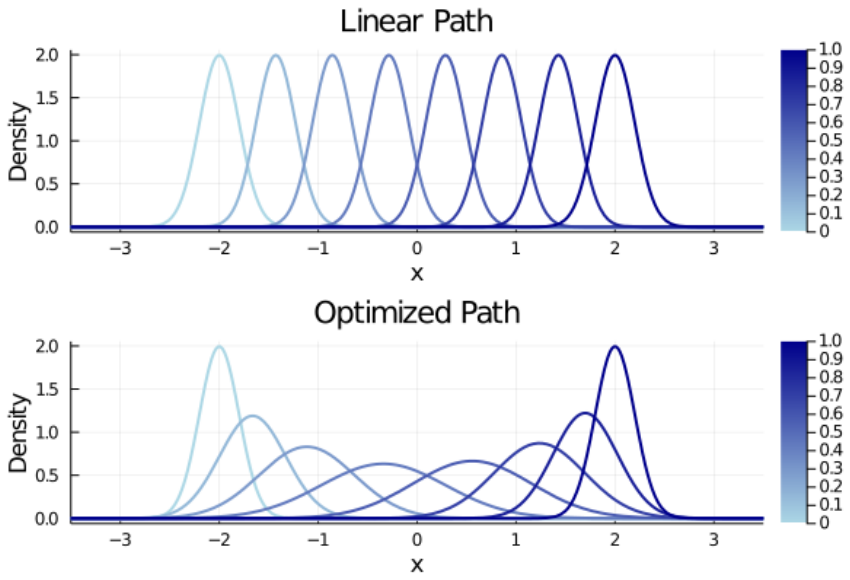

Figure 4.1: Two annealing paths between a $\pi_0 = N(-2, 0.2^2)$ (light blue) and $\pi_1 = N(2, 0.2^2)$ (dark blue) : the traditional linear path (top) and an optimized nonlinear path (bottom). While the distributions in the linear path are nearly mutually singular, those in the optimized path overlap substantially, leading to faster round trips.

construction and optimization of nonlinear paths for annealed importance sampling type algorithms (Gelman and Meng, 1998; Rischard et al., 2018; Grosse et al., 2013) and recently variational inference (Zimmermann et al., 2021; Masrani et al., 2021; Chen et al., 2021). These algorithms are highly parallel; however, for challenging problems, even when combined with adaptive step size procedures (Zhou et al., 2016) they typically suffer from particle degeneracy in Section 3.4.3. Moreover, these methods use different path optimization criteria which are not well motivated in the context of parallel tempering.

Some special cases of non-linear paths have been used in the PT literature (Whitfield et al., 2002; Tawn et al., 2020). Whitfield et al. (2002) construct a non-linear path inspired by the concept of Tsallis entropy, a generalization of Boltzmann-Gibbs entropy, but do not provide algorithms to optimize over this path family. The work of Tawn et al. (2020), also considers a specific example of a nonlinear path distinct from the ones explored in this paper. However, the construction of the nonlinear path in Tawn et al. (2020) requires knowledge of the location of the modes of π_1 and hence makes their algorithm less broadly applicable than standard PT.

4.2 Parallel tempering on general annealing paths

4.2.1 Annealing paths

Given $\pi_0, \pi_1 \in \mathcal{P}(\mathcal{X})$ we expand the terminology *annealing path* from Section 2.1.1 to describe a continuum of distributions interpolating between π_0 and π_1. Formally, an annealing path between π_0 and π_1 is a one-to-one $\mathcal{P}(\mathcal{X})$-valued function $\beta \mapsto \pi_\beta$, such that (1) for all $x \in \mathcal{X}$, $\pi_\beta(x)$ is continuous in β, and (2) is equal to π_0 and π_1 at β equal to 0 and 1 respectively. We will refer to β as the annealing parameter corresponding to the *annealing distribution* π_β,

$$\pi_\beta(x) = \frac{1}{Z(\beta)} \exp\left(W_\beta(x)\right), \quad x \in \mathcal{X},$$

where $Z(\beta) = \int_{\mathcal{X}} \exp\left(W_\beta(x)\right) \mathrm{d}x < \infty$ is the normalizing constant. We will assume $W_\beta(x)$ can be cheaply evaluated for each $x \in \mathcal{X}$ but not $Z(\beta)$.

The most important example of an annealing path is the linear path $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^\beta$, which was the focus of Chapter 3. However, it does not take much imagination to construct a non-linear path. For example, consider a nonlinear path $\pi_\beta \propto \pi_0^{\eta_0(\beta)}\pi_1^{\eta_1(\beta)}$ where $\eta_i : [0,1] \to \mathbb{R}$ are continuous functions such that $\eta_0(0) = \eta_1(1) = 1$ and $\eta_0(1) = \eta_1(0) = 0$. As long as for all $\beta \in [0,1]$, π_β is a normalizable density this is a valid annealing path between π_0 and π_1. Further, note that the path parameter does not necessarily have to appear as an exponent: consider for example the mixture path $\pi_\beta \propto (1-\beta)\pi_0 + \beta\pi_1$. Section 4.6 provides a more detailed example based on linear splines.

4.2.2 Velocity

An annealing path π is *differentiable* at β with *velocity* $\dot{\pi}_\beta : \mathcal{X} \to \mathbb{R}$ equal to the derivative of the log-likelihood when it exists,

$$\begin{aligned}\dot{\pi}_\beta(x) &:= \frac{\mathrm{d}\log \pi_\beta(x)}{\mathrm{d}\beta} \\ &= \frac{\mathrm{d}W_\beta(x)}{\mathrm{d}\beta} - \frac{\mathrm{d}\log Z(\beta)}{\mathrm{d}\beta}.\end{aligned}$$

The velocity $\dot{\pi}_\beta$ corresponds to the score function of the annealing path π_β, and measures the sensitivity of $\pi_\beta(x)$ to changes in β. Under mild regularity assumptions on $W_\beta(x)$, we have $\dot{\pi}_\beta$ has

mean zero with respect to π_β, and the variance of $\dot{\pi}_\beta$ is the Fisher information of π_β,

$$\mathbb{E}_\beta\left[\dot{\pi}_\beta\right] = 0, \qquad \mathrm{Var}_\beta\left[\dot{\pi}_\beta\right] = \mathrm{Var}_\beta\left[\frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}\right],$$

Notably, the linear path between π_0 and π_1 has velocity,

$$\dot{\pi}_\beta(x) = V(x) + \frac{\mathrm{d}\log Z(\beta)}{\mathrm{d}\beta}, \tag{4.1}$$

where recall $V(x) = W_1(x) - W_0(x)$ is the corresponding log-likelihood ratio between π_0 and π_1 modulo the normalizing constant.

The velocity will be important when analyzing the asymptotic and geometric properties of PT for generalized paths. From now on we will assume all annealing paths are continuously piece-wise differentiable.

4.2.3 Path reparametrization

Suppose π_β is a differentiable annealing path with schedule $\mathcal{B}_N$ generated by γ for some increasing differentiable γ with derivative $\dot{\gamma}$. Then $\pi_{\beta_n} = \pi'_{n/N}$ where $\pi'_w = \pi_{\gamma(w)}$ is a differentiable annealing path satisfying,

$$\begin{aligned} \pi'_w &= \pi_{\gamma(w)}, \\ \dot{\pi}'_w &= \dot{\pi}_{\gamma(w)}\dot{\gamma}(w). \end{aligned} \tag{4.2}$$

Choosing a schedule $\mathcal{B}_N$ generated by γ for π is equivalent to the uniform schedule for π'. Consequentially, schedule generators can be equivalently reinterpreted as orientation preserving reparametrizations of an annealing path.

We will say that two differentiable annealing paths π and π' are equivalent if and only if there is a continuous increasing $\gamma : [0,1] \to [0,1]$ such that $\gamma(0) = 0, \gamma(1) = 1$ such that (4.2) holds. Equivalent paths share the same annealing distributions in $\mathcal{P}(\mathcal{X})$ and end points, although they differ in their velocity.

4.2.4 Non-asymptotic analysis

Notice that the local exploration and communication kernels $\mathbf{K}^{\text{expl}}$ and $\mathbf{K}_t^{\text{comm}}$ constructed in Sections 2.2.1 and 2.2.2 respectively are well-defined for general annealing paths π and schedule $\mathcal{B}_N$. In particular, note that Algorithms 1 and 2 hold for any choice of annealing path and do not require the use of a linear path. So we can therefore extend the notion of the index process and round trip to general annealing paths. We will denote $\tau(\pi, \mathcal{B}_N)$ as the round trip rate for a general path π with schedule $\mathcal{B}_N$.

Notably for general annealing paths $\Delta W_n = W_{\beta_n} - W_{\beta_{n-1}}$ may not simplify to (3.1) as it did for the linear path. We only used the structure of the linear path when developing the non-asymptotic theory for PT in Section 3.1 through the ELE assumption (A2) from Section 3.1.2. Assumptions (A1) and (A2) can be simply modified for general paths:

(A1') *Stationarity*: $\mathbf{X}_0 \sim \boldsymbol{\pi}$ and thus $\mathbf{X}_t \sim \boldsymbol{\pi}$ for all t as the kernel $\mathbf{K}_t^{\text{PT}}$ is $\boldsymbol{\pi}$-invariant.

(A2') *Efficient Local Exploration (ELE)*: If $\mathbf{X} \sim \boldsymbol{\pi}$ and $\bar{\mathbf{X}}|\mathbf{X} \sim \mathbf{K}^{\text{expl}}(\mathbf{X}, \mathrm{d}\bar{\boldsymbol{x}})$, then $\Delta W_n(X^{n-1})$ is independent of $\Delta W_n(\bar{X}^{n-1})$ and $\Delta W_n(X^n)$ is independent of $\Delta W_n(\bar{X}^n)$ for all $n = 1, \ldots, N$.

When π is the linear path (A2') reduce to (A2) since $W_n(x) \propto V(x)$. Recall $\alpha^{(n-1,n)}(\mathbf{X}_t)$ is the probability a swap occurs between chains X_t^{n-1} and X_t^n at scan t defined by

$$\alpha^{(n-1,n)}(\boldsymbol{x}) = 1 \wedge \exp(\Delta W_n(x^{n-1}) - \Delta W_n(x^n)).$$

Assumptions (A1') and (A2') ensure for each $n = 1, \ldots, N$ we have $\{\alpha^{(n-1,n)}(\mathbf{X}_t)\}_{t=1}^{\infty}$ are independent and Theorem 1 and Corollary 2 still hold for general annealing paths π with schedule $\mathcal{B}_N$. In particular, the round trip rate satisfies,

$$\tau(\pi, \mathcal{B}_N) = \frac{1}{2 + 2\Lambda(\pi, \mathcal{B}_N)}, \tag{4.3}$$

where $\Lambda(\pi, \mathcal{B}_N)$ is the non-asymptotic communication barrier for path π_β with schedule $\mathcal{B}_N$ defined by,

$$\Lambda(\pi, \mathcal{B}_N) = \sum_{n=1}^{N} \frac{r(\pi_{\beta_{n-1}}, \pi_{\beta_n})}{1 - r(\pi_{\beta_{n-1}}, \pi_{\beta_n})}.$$

Recall $r(\pi_{\beta_{n-1}}, \pi_{\beta_n})$ is the average rejection rate for the n-th swap kernel from Section 3.2.1.

4.3 Asymptotic analysis

Our next objective is to characterize the asymptotic efficiency of a nonlinear path π with schedule $\mathcal{B}_N$ in the regime where $N \to \infty$ —which establishes its fundamental ability to take advantage of parallel computation. In other words, we require a generalization of the asymptotic result from Corollary 5 and Theorem 6. We will follow a similar approach to Section 3.2 in Chapter 3 from the linear path to a general annealing path. In particular, in Chapter 3, we developed the asymptotic theory of parallel tempering for the linear annealing path through the communication barrier Λ. In this section, we will extend the communication barrier for sufficiently regular annealing paths.

4.3.1 Regular annealing paths

We will say that that an annealing path $\pi_\beta \propto \exp(W_\beta)$ is *regular* if for all $x \in \mathcal{X}$, $W_\beta(x)$ is piecewise twice continuously differentiable in β, and when $\frac{\mathrm{d}W}{\mathrm{d}\beta}(x)$ and $\frac{d^2W}{d\beta^2}(x)$ exist and there exists some functions $V_1, V_2 : \mathcal{X} \to [0, \infty)$ such that

$$\forall x \in \mathcal{X}, \ \sup_{\beta \in [0,1]} \left| \frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(x) \right| \leq V_1(x), \tag{4.4}$$

$$\forall x \in \mathcal{X}, \ \sup_{\beta \in [0,1]} \left| \frac{d^2W_\beta}{d\beta^2}(x) \right| \leq V_2(x). \tag{4.5}$$

Moreover, we assume there exists some $\epsilon > 0$, satisfying

$$\sup_\beta \mathbb{E}_\beta[(1 + V_1^3) \exp(\epsilon V_2)] < \infty. \tag{4.6}$$

The differentiability condition (4.4), (4.5) ensures the log-likelihood does not change too rapidly for adjacent distributions along the path. The integrability condition (4.6) is required to control the tail behaviour of distributions formed by linearized approximations to the path π_β. This condition is satisfied for any reasonable path families that would arise in practice. For example, e.g., if $W_\beta(x)$ is piecewise linear for all $x \in \mathcal{X}$, then $V_2 = 0$ trivially.

4.3.2 Communication barrier for regular paths

If $p \in \mathcal{P}(\mathcal{X})$ and f is integrable with respect to p, we define $\lambda(p, f)$ as,

$$\lambda(p, f) := \frac{1}{2}\mathbb{E}_p\left[\left|f(X) - f(X')\right|\right], \quad X, X' \overset{\text{i.i.d.}}{\sim} p. \tag{4.7}$$

Recall from equation (4.1), the linear path π_β between π_0 and π_1 has velocity $\dot{\pi}(x) = V(x) - \frac{\mathrm{d}\log Z}{\mathrm{d}\beta}$. By substituting $V(x) = \dot{\pi}(x) + \frac{\mathrm{d}\log Z}{\mathrm{d}\beta}$ into (3.13) we can rewrite the communication barrier Λ as,

$$\Lambda = \int_0^1 \lambda(\pi_\beta, \dot{\pi}_\beta)\mathrm{d}\beta. \tag{4.8}$$

Notice that (4.8) is well-defined when π is any differentiable annealing path and motivates the definition of $\Lambda(\pi)$,

$$\lambda(\beta) := \lambda(\pi_\beta, \dot{\pi}_\beta), \qquad \Lambda(\pi) := \int_0^1 \lambda(\beta)\mathrm{d}\beta. \tag{4.9}$$

Theorem 10 shows that for regular annealing paths, $\lambda(\beta)$ and $\Lambda(\pi)$ defined in (4.9) encode the same asymptotic information for PT as the local and global communication barrier for the linear path. In particular, Theorem 10 extends Theorem 4, Corollary 5 and Theorem 6 for general non-linear paths, with slightly weaker error estimates. Consequentially the theoretical analysis and methodology developed for the linear path in Chapter 3 are still valid for any regular paths. Motivated by Theorem 10, we will now refer to $\lambda(\beta)$, and $\Lambda(\pi)$ defined through (4.9), as the local and global communication barrier for π.

Theorem 10. *Suppose π is a regular annealing path, with local communication barrier $\lambda(\beta)$ and global communication barrier $\Lambda(\pi)$ defined by* (4.9)*, then we have the following holds:*

(a) The local communication barrier equals the instantaneous rate of rejection,

$$\lim_{\Delta\beta \to 0} \frac{r(\pi_{\beta+\Delta\beta}, \pi_\beta)}{|\Delta\beta|} = \lambda(\beta). \tag{4.10}$$

(b) The total rejection rate uniformly converges to the global communication barrier as $\|\mathcal{B}_N\| \to 0$,

$$\lim_{\delta \to 0} \sup_{\mathcal{B}_N : \|\mathcal{B}_N\| \leq \delta} \left| \sum_{n=1}^{N} r(\pi_{\beta_{n-1}}, \pi_{\beta_n}) - \Lambda(\pi) \right| = 0. \tag{4.11}$$

(c) The asymptotic round rate for non-reversible PT uniformly converges to $\tau_\infty(\pi) := (2+2\Lambda(\pi))^{-1}$ *as* $\|\mathcal{B}_N\| \to 0$,

$$\lim_{\delta \to 0} \sup_{\mathcal{B}_N : \|\mathcal{B}_N\| \leq \delta} |\tau(\pi, \mathcal{B}_N) - \tau_\infty(\pi)| = 0. \tag{4.12}$$

See Appendix A.2.1 for the proof.

Invariance of the communication barrier

Suppose π' is equivalent to π with local communication barrier λ' and λ respectively. There is an orientation preserving reparametrization γ such that (4.2) holds, which we can substitute into (4.7) to get the following relation,

$$\lambda'(w) = \lambda(\gamma(w))\dot{\gamma}(w). \tag{4.13}$$

By integrating (4.13) over w and subtituting $\beta = \gamma(w)$ we have

$$\Lambda(\pi') = \int_0^1 \lambda(\gamma(w))\dot{\gamma}(w)\mathrm{d}w = \int_0^1 \lambda(\beta)\mathrm{d}\beta = \Lambda(\pi).$$

Therefore the global communication barrier Λ is invariant to orientation preserving reparameterizations and cannot distinguish between equivalent paths.

4.4 Path tuning

4.4.1 Annealing path families

It is often the case that there are a set of candidate annealing paths in consideration for a particular target π_1. For example, if a path has tunable parameters $\phi \in \Phi$ that govern its shape, we can generate a collection of annealing paths (up to equivalency) that all target π_1 by varying the

parameter ϕ. We call such collections an annealing path family. Formally, an *annealing path family* for target π_1 is a collection of regular annealing paths $\mathcal{A} = \{\pi^\phi : \phi \in \Phi\}$ such that for all parameters $\phi \in \Phi$, $\pi_1^\phi = \pi_1$.

There are many ways to construct useful annealing path families. For example, if one is provided a parametric family of variational distributions $\{q_\phi : \phi \in \Phi\}$ for some parameter space Φ, one can construct the annealing path family $\mathcal{A} = \{\pi^\phi : \phi \in \Phi\}$ of linear paths $\pi_\beta^\phi = q_\phi^{1-\beta}\pi_1^\beta$ from a variational reference q_ϕ to the target π_1. More generally, given $\eta_i(\beta)$ satisfying the constraints in Section 4.2.1, $\pi_\beta^\phi = q_\phi^{\eta_0(\beta)}\pi_1^{\eta_1(\beta)}$ defines a nonlinear annealing path family. Another example of an annealing path family used in the context of PT are q-paths $\{\pi_\beta^q\}_{q\in[0,1]}$ (Whitfield et al., 2002). Given a fixed reference and target π_0, π_1, the path π_β^q interpolates between the mixture path ($q = 0$) and the linear path ($q = 1$) (Brekelmans et al., 2020). In Section 4.6, we provide a new flexible class of nonlinear paths based on splines that is designed specifically to enhance the performance of PT.

4.4.2 Optimizing over annealing path families

Motivated by the analysis of Section 4.3, given an annealing path family $\mathcal{A} = \{\pi^\phi : \phi \in \Phi\}$, a natural objective function for this optimization to consider is to maximize non-asymptotic round trip rate $\tau(\pi^\phi, \mathcal{B}_N)$ or equivalently minimize the non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ which now depend both on the schedule and path parameter, denoted by superscript ϕ. Since every path in $\mathcal{A}$ has the desired target distribution π_1, we are free to optimize the path over the tuning parameter space $\phi \in \Phi$ in addition to optimizing the schedule $\mathcal{B}_N$,

$$\begin{aligned} \phi^\star, \mathcal{B}_N^\star &= \arg\min_{\phi\in\Phi,\mathcal{B}_N} \Lambda(\pi^\phi, \mathcal{B}_N), \\ &= \arg\min_{\phi\in\Phi,\mathcal{B}_N} \sum_{n=1}^{N} \frac{r_n^\phi}{1 - r_n^\phi}, \end{aligned} \tag{4.14}$$

where $r_n^\phi = r(\pi_{\beta_{n-1}}^\phi, \pi_{\beta_n}^\phi)$. We solve this optimization using an approximate coordinate-descent procedure, iterating between an update of the schedule $\mathcal{B}_N$ for a fixed path parameter $\phi \in \Phi$, followed by a gradient step in ϕ based on a surrogate objective function and a fixed schedule. We assume that the optimization over ϕ ends after a finite number of adaptive rounds to sidestep the potential pitfalls of adaptive MCMC methods (Andrieu and Moulines, 2006). This is summarized in

Algorithm 6 PathOptimizedNRPT

Input: Initial state $\boldsymbol{x}_0$, annealing path family $\{\pi_\beta^\phi : \phi \in \Phi\}$, maximum chains $\bar{N}$, tuning budget t_{tune}, scans per round t_{round}, learning rate κ

▷ Initialize state

1: $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$

▷ Initialize schedule, e.g. uniform

2: $\mathcal{B}_{\bar{N}} \leftarrow (0, 1/\bar{N}, 2/\bar{N}, \ldots, 1)$

▷ Initialize linear path, e.g. linear path

3: $\pi_\beta^\phi \leftarrow \pi_0^{1-\beta}\pi_1^\beta$

4: `maxRound` $\leftarrow t_{\text{tune}}/t_{\text{round}}$

5: **for** `round` **in** 1, 2, ..., `maxRound` **do**

▷ Approximate rejection rate using Algorithm 1

6: $\{\boldsymbol{x}_t\}_{t=1}^{t_{\text{round}}}, (r_n)_{n=1}^{\bar{N}} \leftarrow$ `DEO`$(\boldsymbol{x}, \pi_\beta^\phi, \mathcal{B}_N, t_{\text{round}})$

▷ Approximate communication barrier using Algorithm 4

7: $\lambda, \Lambda \leftarrow$ `CommunicationBarrier`$((r_n)_{n=1}^{\bar{N}}, \mathcal{B}_{\bar{N}})$

▷ Approximate optimal schedule using Algorithm 3

8: $\mathcal{B}_{\bar{N}} \leftarrow$ `UpdateSchedule`$(\lambda, \bar{N})$

▷ Update path using gradient descent (See Section 4.4.4)

9: $\phi \leftarrow \phi - \kappa\nabla_\phi \text{SKL}(\pi^\phi, \mathcal{B}_{\bar{N}})$

10: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{t_{\text{scan}}}$

11: **end for**

12: **return** $\pi_\beta^\phi, \mathcal{B}_{\bar{N}}, \lambda, \Lambda$

Algorithm 6.

Algorithm 6 is not a replacement of Algorithm 5, but rather a companion. Given an annealing path family, we can incorporate path tuning using Algorithm 6 in place of the tuning phase (lines 1-9) of Algorithm 5. To utilize the tuning budget of t_{tune} scans, we run $t_{\text{tune}}/t_{\text{round}}$ adaptive rounds of t_{scan} scans per round instead of the doubling procedure from Algorithm 5. See Syed et al. (2021b) for guidelines on how to tune t_{round} and the learning rate κ.

4.4.3 Optimizing the schedule

Fix the value of ϕ, which fixes the path $\pi^\phi \in \mathcal{A}$ and we will characterize the optimal schedule $\mathcal{B}_N^\phi = \arg\min_{\mathcal{B}_N} \Lambda(\pi^\phi, \mathcal{B}_N)$ for π_β^ϕ by using the same analysis in Section 3.3.1. When N is large, finding the schedule that minimizes $\Lambda(\pi^\phi, \mathcal{B}_N)$ reduces to approximately solving the following

constraint optimization problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{n=1}^{N} \frac{r_n^\phi}{1 - r_n^\phi}, \\ \text{s.t.} \quad & \sum_{n=1}^{N} r_n^\phi = \Lambda(\pi^\phi). \end{aligned}$$

Based on the same argument in Section 3.3.1, the optimal schedule $\mathcal{B}_N^\phi = (\beta_n^\phi)_{n=1}^N$ achieves a constant rejection $r_n^\phi = r^\phi$ for $n = 1, \ldots, N$ and approximately satisfies,

$$\frac{1}{\Lambda(\pi^\phi)} \int_0^{\beta_n^\phi} \lambda^\phi(\beta) \mathrm{d}\beta = \frac{n}{N}, \tag{4.15}$$

where λ^ϕ is the local communication barrier for π^ϕ. We want to remark that condition (4.15) implies $\beta_n^\phi = \gamma^\phi(n/N)$, where γ^ϕ is the inverse of $F^\phi(\beta) = \frac{1}{\Lambda^\phi} \int_0^\beta \lambda^\phi(u) \mathrm{d}u$. Consequentially, the optimal schedule $\mathcal{B}_N^\phi$ for the path π_β^ϕ is generated by γ^ϕ.

Given π^ϕ and the current schedule $\mathcal{B}_N = (\beta_n)_{n=0}^N$, we can run PT and obtain Monte Carlo estimates for r_n^ϕ. We can then use Algorithm 4 to efficiently approximate the communication barrier for π^ϕ using monotone splines. Equipped with the communication barrier, Algorithm 3 can be used to solve (4.15) and obtain an estimate of $\mathcal{B}_N^\phi$.

4.4.4 Optimizing the path

Fix the schedule $\mathcal{B}_N$; we now want to improve the path π^ϕ itself by modifying ϕ. However, in challenging problems, this is not as simple as taking a gradient step for the non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ objective in Equation (4.14). In particular, in early iterations—when the path is near its oft-poor initialization—the rejection rates satisfy $r_n^\phi \approx 1$. As demonstrated empirically in (Syed et al., 2021b, Appendix F), gradient estimates in this regime exhibit a low signal-to-noise ratio that precludes their use for optimization. Moreover, the non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ depends on the path π^ϕ, and its parameterization through the annealing schedule $\mathcal{B}_N$. Since we want to optimize the path independent of the parameterization, we will construct a proxy objective for $\Lambda(\pi^\phi, \mathcal{B}_N)$ that (1) measures the quality of the path, (2) is numerically stable to optimize.

Using the convexity of $r \to r/(1-r)$, the non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ satisfies,

$$\Lambda(\pi^\phi, \mathcal{B}_N) = \sum_{n=1}^{N} \frac{r_n^\phi}{1 - r_n^\phi} \geq \frac{r(\pi^\phi, \mathcal{B}_N)}{1 - N^{-1} r(\pi^\phi, \mathcal{B}_N)}, \tag{4.16}$$

where $r(\pi^\phi, \mathcal{B}_N) := \sum_{n=1}^{N} r_n^\phi$ is the total rejection rate, and we have equality if and only if $\mathcal{B}_N$ achieves equi-acceptance. We can equivalently minimize $r(\pi^\phi, \mathcal{B}_N)$ since the lower bound in (4.16) is approximately attained when the optimal schedule $\mathcal{B}_N^\phi$ for π^ϕ is chosen and r_n^ϕ is constant $n = 1, \ldots, N$.

Theorem 10 implies the total rejection rate $r(\pi^\phi, \mathcal{B}_N) \approx \Lambda(\pi^\phi)$ when $\|\mathcal{B}_N\|$ is small. It follows from (4.11) in Theorem 10 that if $(\pi^\phi, \mathcal{B}_N)$ approximately miniminizes $r(\pi^\phi, \mathcal{B}_N)$, then so does $(\pi^\phi, \mathcal{B}'_N)$ for every annealing schedule $\mathcal{B}'_N$ (assuming $\|\mathcal{B}'_N\|$ is small). This means the path minimizing the total rejection rate does not admit a unique solution making it numerically unstable to optimize. To remedy this issue, we will find a suitable surrogate objective that upper bounds $r(\pi^\phi, \mathcal{B}_N)$ and does not suffer from the same identifiability issues.

By Jensen's inequality,

$$r(\pi^\phi, \mathcal{B}_N)^2 = \left(\sum_{n=1}^{N} r_n^\phi \right)^2 \leq N \sum_{n=1}^{N} r^2(\pi^\phi_{\beta_{n-1}}, \pi^\phi_{\beta_n}) \tag{4.17}$$

with equality when the rejection rates are constant. From Equation (3.11) we know that the squared rejection rate is bounded above by the SKL divergence, which combined with (4.17) results in the SKL objective,

$$r(\pi^\phi, \mathcal{B}_N)^2 \leq \mathrm{SKL}(\pi^\phi, \mathcal{B}_N), \tag{4.18}$$

where $\mathrm{SKL}(\pi, \mathcal{B}_N)$ equals,

$$\mathrm{SKL}(\pi, \mathcal{B}_N) := N \sum_{n=1}^{N} \mathrm{SKL}(\pi_{\beta_{n-1}}, \pi_{\beta_n}). \tag{4.19}$$

Since $\mathrm{SKL}(\pi^\phi, \mathcal{B}_N)$ bounds $r(\pi^\phi, \mathcal{B}_N)$, it is a candidate objective to optimize. In the next section we

will argue that optimizing the SKL objective leads to near optimal paths for parallel tempering, and offer theoretical evidence that it does not suffer from the same identifiability issues as $r(\pi^\phi, \mathcal{B}_N)$. See Appendix F of Syed et al. (2021b) for empirical evidence that minimizing the SKL objective leads to improved performance for PT and Appendix D of Syed et al. (2021b) for a derivation of the gradient, $\nabla_\phi \text{SKL}(\pi^\phi, \mathcal{B}_N)$.

The slack in inequality (4.18) could potentially depend on ϕ even in the large N regime. Therefore, during optimization, we recommend monitoring the value of the non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ to ensure that the optimization of the surrogate SKL objective $\text{SKL}(\pi^\phi, \mathcal{B}_N)$ indeed improves it, and hence the round trip rate performance of PT via Equation (4.3). We will explore the asymptotic properties of the SKL objective and characterize its optima for large N in Section 4.5.8.

4.5 Annealing within parametric families

Suppose a regular annealing path π_β takes on values in some parametric family of distributions

$$\mathcal{M} = \{p_\eta \in \mathcal{P}(\mathcal{X}) : \eta \in \Omega\},$$

where Ω is some subset of $\mathbb{R}^d$. The corresponding probability distributions p_η have density,

$$p_\eta(x) = \frac{1}{Z(\eta)} \exp(W_\eta(x)),$$

where $Z(\eta) = \int_{\mathcal{X}} \exp(W_\eta(x)) \mathrm{d}x$ is corresponding normalizing constant, and $W_\eta(x)$ is the log-likelihood modulo a constant. For example suppose $\mathcal{X} = \mathbb{R}$ and $\Omega = \mathbb{R} \times \mathbb{R}_+$, then we can identify each $\eta = (\mu, \sigma) \in \Omega$ with $p_\eta = N(\mu, \sigma^2)$. The corresponding parametric family $\mathcal{M}$ is the space of Gaussians with log-likelihood $W_\eta(x) = \frac{1}{2\sigma^2}(x - \mu)^2$. Annealing paths π in the space of Gaussians can be equivalently represented by the 2-dimensional curve $\eta(\beta) = (\mu(\beta), \sigma(\beta)) \in \Omega$ corresponding to annealing distributions $\pi_\beta = N(\mu(\beta), \sigma(\beta)^2) \in \mathcal{M}$.

More generally, by identifying $p_\eta \in \mathcal{M}$ with $\eta \in \Omega$, the family $\mathcal{M}$ inherits the topological and geometric structure of Ω. In particular, we can study the geometric properties of annealing paths π_β in $\mathcal{M}$ through the geometry of d-dimensional curves $\eta(\beta)$ in the space of parameters Ω. With

this interpretation, the local and global communication barriers will emerge as the "speed" and "length" of annealing paths, respectively and provide an alternative view of Algorithm 6.

4.5.1 Annealing families

Given a parameteric family $\mathcal{M}$, we define the *score function* for $\mathcal{M}$ as

$$S_\eta(x) := \nabla_\eta \log p_\eta(x),$$

and the *Fisher information* matrix $I(\eta)$,

$$I(\eta) := -\mathbb{E}_\eta[\nabla_\eta^2 \log p_\eta],$$

where we use $\mathbb{E}_\eta$ to indicate the expectation is with respect to p_η. We will say that a parametric family $\mathcal{M}$ is a *annealing family* if the following conditions hold:

(R1) *Identifiability:* For all $\eta, \eta' \in \Omega$ if $\eta' \neq \eta$, then $p_\eta \neq p_{\eta'}$.

(R2) *Topological regularity:* Ω is complete, locally compact and path connected.

(R3) *Geometric regularity:* $W_\eta(x)$ is twice continuously differentiable in η for all x and for all compact $K \subset \Omega$, there exist functions $V_1, V_2 : \mathcal{X} \to \mathbb{R}$ such that

$$\forall x \in \mathcal{X}, \sup_{\eta \in K,\ \|v\|=1} \left| v^T \nabla_\eta W_\eta(x) \right| \leq V_1(x),$$
$$\forall x \in \mathcal{X}, \sup_{\eta \in K,\ \|v\|=1} \left| v^T \nabla_\eta^2 W_\eta(x) v \right| \leq V_2(x),$$

for some functions $V_1, V_2 : \mathcal{X} \to [0, \infty)$ and there exists some $\epsilon > 0$ satisfying

$$\sup_{\eta \in K} \mathbb{E}_\eta[(1 + V_1^3) \exp(\epsilon V_2)] < \infty. \tag{4.20}$$

(R4) *Positive definite:* The Fisher information is positive definite for all $\eta \in \Omega$,

$$\eta \in \Omega, v \neq 0, \qquad v^T I(\eta) v > 0.$$

Condition (R1) is there to make sure the parametric representation of $\mathcal{M}$ is uniquely determined by $\eta \in \Omega$. (R2) and (R3) ensure $\mathcal{M}$ is compatible with parallel tempering. In particular, condition (R2) guarantees the existence of annealing paths between any two points $\pi_0, \pi_1 \in \mathcal{M}$, and rules pathologies arising from multiple connected components for $\mathcal{M}$. (R3) ensures that any piecewise twice-differentiable annealing paths in $\mathcal{M}$ are regular and thus satisfy the conditions of Theorem 10. Moreover, (R3) also guarantees we can twice differentiate $\int_{\mathcal{X}} p_\eta(x)\mathrm{d}x$ under the integral sign with respect to η which implies the score function has mean 0 with respect to η, and the Fisher information is the covariance matrix for the score function, i.e. for all $\eta \in \Omega$,

$$\mathbb{E}_\eta[S_\eta] = 0, \qquad \mathbb{E}_\eta[S_\eta S_\eta^T] = I(\eta). \tag{4.21}$$

Finally (R4) makes sure that $I(\eta)$ is always full rank and eliminates the possibility that Ω is a low dimensional surface embedded in $\mathbb{R}^d$.

In particular, the image of any regular annealing path π satisfying the condition of Theorem 10 is a 1-dimensional annealing family $\mathcal{M} = \{\pi_\beta : \beta \in [0,1]\}$ parametrized by $\beta \in [0,1]$ with Fisher information $I(\beta) = \mathrm{Var}_\beta\left[\frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}\right]$. This is not an interesting family, since there only exists one regular paths between π_0 and π_1 up to equivalency, namely π, but demonstrates the generality of our framework. We will see that even a 2-dimensional annealing family $\mathcal{M}$ will give us enough freedom to design flexible annealing path families.

Tangent bundle

For an annealing family $\mathcal{M}$, it is possible to fully characterize all differentiable annealing paths in terms of the corresponding differentiable curves in Ω. Define *tangent bundle* of $\mathcal{M}$ denoted $T\mathcal{M}$ as the possible values for a differentiable annealing path in $\mathcal{M}$ and its velocity,

$$T\mathcal{M} = \{(\pi_\beta, \dot{\pi}_\beta) : \pi \text{ is a differentible annealing path in } \mathcal{M}\}.$$

Proposition 11 implies that $T\mathcal{M}$ is in bijective correspondence to $\Omega \times \mathbb{R}^d$, and therefore there is an equivalence between differentiable paths π in $\mathcal{M}$ and differentiable parametric curves η in Ω.

Proposition 11. *Suppose $\mathcal{M}$ is an annealing family, then $\Phi : \Omega \times \mathbb{R}^d \to T\mathcal{M}$ defined by*

$$\Phi(\eta, v) = (p_\eta, v^T S_\eta),$$

is a bijection.

See Appendix A.2.2 for the proof.

4.5.2 Regular divergences on annealing families

A *divergence* on an annealing family $\mathcal{M}$ is a function $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ such that for all $p, p' \in \mathcal{M}$:

$$D(p, p') \geq 0, \qquad D(p, p') = 0 \iff p = p'. \tag{4.22}$$

Intuitively, $D(p, p)$ measures the difference between p and p', but may not be symmetric or satisfy the triangle inequality. Divergences are helpful since they provide a more flexible notion of "distance" between probability distributions than a metric. In particular, the divergences relevant for PT are the rejection rate and SKL divergence defined in Section 3.2.1.

Since a divergence can be general, to say anything meaningful we require some structure on D to ensure $D(p_\eta, p_{\eta'})$ is well behaved when $\eta \approx \eta'$. If $D(p_\eta, p'_\eta)$ is sufficiently smooth as a function of η, η', then (4.22) implies that at $D(p_\eta, p_{\eta'})$ and its partial derivatives should vanish at $\eta = \eta'$ and the Hessian should be positive definite. Then by Taylor's Theorem we expect $D(p_\eta, p_{\eta+\Delta\eta})$ to be locally quadratic in $\Delta\eta$ for small perturbations $\Delta\eta$ from η. Formally, we will say D is a *regular divergence* if for each $\eta \in \Omega$, as $\|\Delta\eta\| \to 0$,

$$D(p_\eta, p_{\eta+\Delta\eta}) = \lambda_D^2(\eta, \Delta\eta) + o(\|\Delta\eta\|^2), \tag{4.23}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$ and $\lambda_D : \Omega \times \mathbb{R}^d \to \mathbb{R}_+$ satisfies,

(F1) *Regularity:* $\lambda_D(\eta, v)$ is continuously differentiable on $\{(\eta, v) \in \Omega \times \mathbb{R}^d : v \neq 0\}$.

(F2) *Positive definite:* $\lambda_D(\eta, v) \geq 0$ with equality if and only if $v = 0$.

(F3) *Positive homogeneity:* For all $c \geq 0$, $\lambda_D(\eta, cv) = c\lambda_D(\eta, v)$.

(F4) *Strong convexity:* For all $\eta \in \Omega$, $v \to \lambda_D^2(\eta, v)$ is strongly convex, i.e. for all $v \neq 0$,

$$G_D(\eta, v) = \nabla_v^2 \lambda_D^2(\eta, v) \succ 0.$$

We call $\lambda_D(\eta, v)$ and $G_D(\eta, v)$ the *Finsler norm* and *Finsler metric* induced by divergence D. When $\mathcal{M}$ is an annealing family, we can use Proposition 11 to define the Finsler norm and metric on the tangent bundle $T\mathcal{M}$,

$$\lambda_D(p_\eta, v^T S_\eta) := \lambda_D(\eta, v), \qquad G_D(p_\eta, v^T S_\eta) := G_D(\eta, v). \tag{4.24}$$

The Finsler norm $\lambda_D(\eta, \Delta\eta)$ characterizes the local behaviour of the divergence. Condition (F1) ensures $\lambda_D(\eta, v)$ smoothly varies and does not undergo sudden jumps. Equation (4.23) combined with conditions (F2)-(F3) ensure that $D(p_\eta, p_{\eta+\Delta\eta})$ is locally quadratic in $\Delta\eta$ for all η. Finally, (F4) ensures $D(\eta, \eta')$ is uniquely minimized at $\eta = \eta'$. It can sometimes be easier to verify the strict sub-additivity condition (F4') which is equivalent to (F4) (Tamássy, 2005),

(F4') *Strict sub-additivity:* For all $\eta \in \Omega$, and $v' \neq cv$ for some $c \neq 0$,

$$\lambda_D(\eta, v + v') < \lambda_D(\eta, v) + \lambda_D(\eta, v').$$

Intuitively $\lambda_D(\eta, v)$ is an asymmetric norm for v that smoothly varies with η. Notice that (F3) does not imply that $\lambda_D(\eta, v)$ is symmetric in v since it is possible that $\lambda_D(\eta, -v) \neq \lambda_D(\eta, v)$ accounting for the potential asymmetry of the divergence. When $\lambda_D(\eta, -v) = \lambda_D(\eta, v)$, the Finsler norm reduces to a norm for each $\eta \in \Omega$, and thus (4.23) implies a regular divergence locally behaves like a square norm.

4.5.3 Regularity of f-divergences

When $D(p_\eta, p_{\eta'})$ is twice-continuously differential in η, η', then it follows by Taylor's theorem,

$$D(p_\eta, p_{\eta+\Delta\eta}) = \Delta\eta^T G_D(\eta) \Delta\eta + o(\|\Delta\eta\|^2),$$

where $G_D(\eta)$ is proportional to the Hessian of $\eta' \mapsto D(\eta, \eta')$ evaluated at $\eta' = \eta$. In this case the Finsler norm is induced by an inner product and $\lambda^2(\eta, v) = v^T G_D(\eta) v$ is a quadratic form for each $\eta \in \Omega$. Therefore, the Finsler metric induced by D is constant in v, and $G_D(\eta, v) = G_D(\eta)$ coincides with a *Riemannian metric*.

Given a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$, define the f-divergence on $\mathcal{P}(\mathcal{X})$ as

$$D_f(p, p') = \int_{\mathcal{X}} f\left(\frac{p(x)}{p'(x)}\right) p(x) \mathrm{d}x.$$

f-divergences (Csiszár, 1967) are an important family of divergences in statistics and machine learning. For appropriately chosen f, we can recover many commonly used divergences such as the KL, reverse KL, SKL, Chi-square, total variation distance, squared-Hellinger distance, and α-divergences to name a few. It follows from Polyanskiy (2020, Theorem 7.12) that when f is twice differentiable with $\limsup_{u \to \infty} f''(u) < \infty$ then D_f is a regular divergence on $\mathcal{M}$,

$$D_f(p_\eta, p_{\eta + \Delta\eta}) = \frac{f''(1)}{2} \Delta\eta^T I(\eta) \Delta\eta + o(\|\Delta\eta\|^2),$$

with Finsler norm $\lambda_{D_f}(\eta, v) = \sqrt{\frac{f''(1)}{2} v^T I(\eta) v}$, and Finsler metric $G_{D_f}(\eta, v) = \frac{f''(1)}{2} I(\eta)$, where recall $I(\eta)$ is the Fisher information matrix for $\mathcal{M}$.

In particular, $f(u) = \frac{1}{2}(u-1)\log u$, we have $D_f = \mathrm{SKL}$, which implies SKL divergence is regular with Finsler norm and metric,

$$\lambda_{\mathrm{SKL}}(\eta, v) = \sqrt{\frac{1}{2} v^T I(\eta) v}, \qquad G_{\mathrm{SKL}}(\eta, v) = \frac{1}{2} I(\eta).$$

Connection to information geometry

Information geometry is the branch of mathematics devoted to studying the differential-geometrization of statistics. There is a rich literature devoted to studying geometry induced by the Fisher information metric and its applications. See Nielsen (2013) for an elementary introduction to the subject, and Nielsen (2020); Amari (2016) for a more comprehensive overview.

Finsler geometry can be studied independently of divergences and comes equipped with a natural notion of speed, length and distances through the Finsler norm. We refer to Tamássy (2005) for

an overview of the metric geometry induced by the Finsler norm, and Bishop (2013) for a Finsler view of Riemannian geometry. For an in-depth study of the Finsler geometry generated by regular divergences, see Shen (2006).

4.5.4 Regularity of the rejection rate

If a divergence D is not regular, it is often the case that a scaled version is. In general if $D(p_\eta, p_{\eta+\Delta\eta}) \sim C(\eta)\|\Delta\eta\|^\alpha$ as $\Delta\eta \to 0$, then $D^{2/\alpha}(p_\eta, p_{\eta'}) = D(p_\eta, p_{\eta'})^{2/\alpha}$ satisfies the condition of a regular divergence if $C(\eta)$ is sufficiently smooth in η. We will now show that the rejection rate $r(p_\eta, p_{\eta'})$ is not regular, but $r^2(p_\eta, p_\eta)$ is, with the local communication defining a Finsler norm.

Proposition 12. *If $\mathcal{M}$ is an annealing family, then r^2 is regular divergence with Finsler norm, $\lambda_{r^2} : \Omega \times \mathbb{R}^d \to \mathbb{R}_+$*

$$\lambda_{r^2}(\eta, v) = \lambda(p_\eta, v^T S_\eta), \tag{4.25}$$

where $\lambda(p, f)$ is defined in (4.7).

See Appendix A.2.3 for the proof.

4.5.5 Speed and length induced by regular divergences

Suppose D is a regular divergence on an annealing family $\mathcal{M}$, then recall the induced Finsler norm λ_D is well-defined on the tangent bundle $T\mathcal{M}$ through (4.24). If π is a differentiable annealing path in $\mathcal{M}$, we define $\lambda_D(\beta) := \lambda_D(\pi_\beta, \dot{\pi}_\beta)$ and $\Lambda_D(\pi) := \int_0^1 \lambda_D(\beta)\mathrm{d}\beta$ as the *speed* and *length* of π respect to D respectively.

It follows directly from (4.23) that the speed is the derivative of $\sqrt{D}$ along π with respect to β,

$$\lim_{\Delta\beta\to 0} \frac{\sqrt{D(\pi_\beta, \pi_{\beta+\Delta\beta})}}{|\Delta\beta|} = \lambda_D(\beta). \tag{4.26}$$

The speed measures how rapidly π_β changes as measured by D for small changes in β. It can be used to approximate the divergence between nearby points along the path by integrating (4.26),

from β to $\beta + \Delta\beta$,

$$\sqrt{D(\pi_\beta, \pi_{\beta+\Delta\beta})} = \int_\beta^{\beta+\Delta\beta} \lambda_D(u)\mathrm{d}u + o(\Delta\beta). \tag{4.27}$$

Consequentially, for any schedule $\mathcal{B}_N$, by summing (4.27) over $n = 1, \ldots, N$,

$$\sum_{n=1}^{N} \sqrt{D(\pi_{\beta_{n-1}}, \pi_{\beta_n})} = \Lambda_D(\pi) + o(1), \tag{4.28}$$

with equality as $\|\mathcal{B}_N\| \to 0$. This shows that length is the cumulative change along the path measured by D.

Given an annealing path π in $\mathcal{M}$, Proposition 12 allows us to reinterpret the local and global communication barrier as the speed of and length of π with respect to the squared rejection rate,

$$\lambda(\beta) = \lambda_{r^2}(\beta), \qquad \Lambda(\pi) = \Lambda_{r^2}(\pi).$$

More generally, we have shown that (4.10) and (4.11) from Theorem 10 are special instances of (4.26) and (4.28) respectively and hold more generally for regular divergences D on an annealing family $\mathcal{M}$.

Computation of speed and length

Given an annealing schedule $\mathcal{B}_N$ and estimates of $r(\pi_{\beta_{n-1}}, \pi_{\beta_n})$, we can reinterpreted the output of Algorithm 4 as an approximation of the speed and length of π with respect to r^2. More generally (4.27), (4.28) imply Algorithm 4 can be used more broadly to approximate the speed and length of π with respect to any regular divergence D given estimates of $\sqrt{D(\pi_{\beta_{n-1}}, \pi_{\beta_n})}$ rather than $r(\pi_{\beta_{n-1}}, \pi_{\beta_n})$.

4.5.6 Schedule optimization

Reparametrizations

Suppose π' is an equivalent path to π with speed λ'_D and λ_D respectively. Then there is an orientation preserving γ such that $\pi'_w = \pi_{\gamma(w)}$. Using condition (F3) and chain rule,

$$\lambda'_D(w) = \lambda_D(\gamma(w))\dot{\gamma}(w). \tag{4.29}$$

By substituting $\beta = \gamma(w)$ and integrating (4.29), it follows that all equivalent annealing paths have the same length with respect to D:

$$\Lambda_D(\pi') = \int_0^1 \lambda_D(\gamma(w))\dot{\gamma}(w)\mathrm{d}w = \int_0^1 \lambda_D(\beta)\mathrm{d}\beta = \Lambda_D(\pi).$$

Since the length is invariant to reparametrization, it begs the question: is there a natural reparameterization for π compatible with D? The answer is yes, and they correspond to the constant speed reparametrization.

Paths of constant speed

Given an annealing path π, we will say that γ is a constant speed parametrization for π if $\pi' = \pi_{\gamma(w)}$ has constant speed, i.e. $\lambda'_D(w)$ is constant for all w. Since the length of π is invariant to reparametrization, for all $w \in [0, 1]$, the speed must equal the length,

$$\lambda'_D(w) = \Lambda(\pi). \tag{4.30}$$

It follows immediately from (4.27) that π' is constant speed if and only if the schedule generated by γ satisfies

$$\sqrt{D\left(\pi_{\beta_{n-1}}, \pi_{\beta_n}\right)} = \frac{\Lambda_D(\pi)}{N} + o(N^{-1}),$$

with equality as $N \to \infty$. This implies that the schedule $\mathcal{B}_N$ generated by γ, will ensure $D(\pi_{\beta_{n-1}}, \pi_{\beta_n})$ is approximately constant for all $n = 1, \ldots, N$.

Computation of optimal schedule

When $D = r^2$, we can reinterpret the optimal schedule generator derived in Section 3.3 and 4.4.3 as constant speed reparameterization of π with respect to the rejection rate. More generally, given π with speed $\lambda_D(\beta)$ we can directly compute γ by substituting (4.29) into (4.30), and integrating both sides

$$\int_0^w \lambda_D(\gamma(u))\dot{\gamma}(u)\mathrm{d}u = \Lambda_D(\pi)w. \tag{4.31}$$

After substituting in $\beta = \gamma(u)$ into (4.31),

$$\int_0^{\gamma(w)} \lambda_D(\beta)\mathrm{d}\beta = \Lambda_D(\pi)w.$$

Therefore $\gamma(w) = F^{-1}(w)$ where $F(\beta) = \frac{1}{\Lambda_D(\pi)} \int_0^\beta \lambda_D(u)\mathrm{d}u$. If we have estimates of the speed $\lambda_D(\beta)$ and length $\Lambda_D(\pi)$, then we can use Algorithm 3 to compute the schedule generator to achieve an approximately constant value of $D(\pi_{\beta_{n-1}}, \pi_{\beta_n})$ for all $n = 1, \ldots, N$.

4.5.7 Geodesics

Equipped with the length, we can define the *distance* between points in $\mathcal{M}$ as the length of the shortest path with respect to D. Given $\pi_0, \pi_1 \in \mathcal{M}$, we define

$$d_D(\pi_0, \pi_1) = \inf_{\pi \in \mathcal{A}(\pi_0, \pi_1)} \Lambda_D(\pi),$$

where $\mathcal{A}(\pi_0, \pi_1)$ is the set of regular paths π in $\mathcal{M}$ with endpoints π_0 and π_1. In general d_D is a semi-metric (positive definite, satisfies the triangle inequality) that generates the same topology for $\mathcal{M}$ induced by open sets in Ω, and is locally compatible with the regular divergence D, i.e.

$$\lim_{\eta \to \eta'} \frac{d_D(p_\eta, p_{\eta'})^2}{D(p_\eta, p_{\eta'})} = 1.$$

When $\lambda_D(\eta, -v) = \lambda_D(\eta, v)$, then d_D is also symmetric and hence a metric. See Tamássy (2005, Section 2) for further exposition about the distance function induced by a Finsler norm.

We say an annealing path π in $\mathcal{M}$ is a *geodesic* between π_0, π_1 with respect to D if (1) π is

constant speed and (2) $\Lambda_D(\pi) = d_D(\pi_0, \pi_1)$. The Hofp-Rinow Theorem guarantees the existence of geodesic curve between any two points in $\mathcal{M}$ since (R2) presumes Ω is complete and locally compact (Bridson and Haefliger, 2013, Proposition 3.7). In practice finding geodesics is not an easy task since it requires minimizing the length functional Λ_D. This can be difficult since the minimizer is not unique: in fact if π minimizes Λ_D so does $\pi_{\gamma(w)}$ for every orientation preserving reparameterization γ. This is easily resolved using Jensen's inequality,

$$\Lambda_D(\pi)^2 = \left(\int_0^1 \lambda_D(\beta)\mathrm{d}\beta\right)^2 \leq \int_0^1 \lambda_D^2(\beta)\mathrm{d}\beta,$$

with equality if and only if π is a constant speed path. This implies optimizing $\Lambda_D(\pi)$ is equivalent to minimizing the *kinetic energy* functional

$$E_D(\pi) := \int_0^1 \lambda_D^2(\beta)\mathrm{d}\beta.$$

Unlike the length, the energy $E_D(\pi)$ is not invariant to reparameterization and is minimized by geodesics.

4.5.8 Path optimization

We can now use this geometric structure of the communication barrier to characterize the optimal performance of PT. Equation (4.12) from Theorem 10 implies the optimal round trip rate for a fixed annealing path π in $\mathcal{M}$ and all possible annealing schedule is,

$$\tau(\pi) = \frac{1}{2 + 2\Lambda_{r^2}(\pi)}.$$

By taking the supremum over all annealing paths between reference π_0 and target π_1 in $\mathcal{M}$, then the theoretically optimal round trip rate for PT when annealing in $\mathcal{M}$ is,

$$\sup_{\pi \in \mathcal{A}(\pi_0, \pi_1)} \tau(\pi) = \frac{1}{2 + 2d_{r^2}(\pi_0, \pi_1)}.$$

Therefore $d_{r^2}(\pi_0, \pi_1)$ encodes the fundamental limit of parallel tempering among all possible annealing paths on $\mathcal{M}$, optimally chosen schedule $\mathcal{B}_N$, and an infinite number of parallel chains N.

Given an annealing path π_β, It also follows immediately from Jensen's inequality, that

$$\lambda(\pi_\beta, \dot{\pi}_\beta) \leq \lambda_{\mathrm{SKL}}(\pi_\beta, \dot{\pi}_\beta), \tag{4.32}$$

and thus $\Lambda(\pi) \leq \Lambda_{\mathrm{SKL}}(\pi)$, where $\Lambda_{\mathrm{SKL}}(\pi)$ is proportional to the Fisher-Rao length of π,

$$\sup_{\pi \in \mathcal{A}(\pi_0, \pi_1)} \sup_{\mathcal{B}_N} \tau(\pi, \mathcal{B}_N) \geq \frac{1}{2 + 2d_{\mathrm{SKL}}(\pi_0, \pi_1)}. \tag{4.33}$$

Proposition 7 shows that in the high dimensional scaling limit, the local and global communication barriers are asymptotically equivalent to the speed and length with respect to the Fisher information,

$$\begin{aligned}
\lambda(\beta) &\sim \sqrt{\frac{2}{\hat{\pi}}}\, \lambda_{\mathrm{SKL}}(\beta) \approx 0.798 \lambda_{\mathrm{SKL}}(\beta), \\
\Lambda(\pi) &\sim \sqrt{\frac{2}{\hat{\pi}}}\, \Lambda_{\mathrm{SKL}}(\pi) \approx 0.798 \Lambda_{\mathrm{SKL}}(\pi),
\end{aligned}$$

where $\hat{\pi}$ is the constant $3.1415\ldots$. This offers evidence that in the high dimensional regime the slack in (4.32) and (4.33) is independent of the path, and the geodesics with respect to the SKL are approximately geodesics with respect to r^2 and thus still provide near optimal performance for PT.

Since π is a regular path, with schedule $\mathcal{B}_N$ generated by γ, Theorem 1 from Grosse et al. (2013) implies as $N \to \infty$, the SKL objective $\mathrm{SKL}(\pi, \mathcal{B}_N)$ defined in (4.19) converges to the kinetic energy of the reparameterized path $\pi'_w = \pi_{\gamma(w)}$.

Proposition 13. *If π is a regular path on an annealing family $\mathcal{M}$ and $\mathcal{B}_N$ is generated by γ, then*

$$\lim_{N \to \infty} \mathrm{SKL}(\pi, \mathcal{B}_N) = E_{\mathrm{SKL}}(\pi'),$$

where $\pi'_w = \pi_{\gamma(w)}$.

This implies for large N minimizing the SKL objective is equivalent to minimizing the energy functional, which is minimized at the geodesic. Given an annealing path family $\mathcal{A}$, the path $\pi^* \in \arg\min_{\pi \in \mathcal{A}} \mathrm{SKL}(\pi, \mathcal{B}_N)$ can be interpreted as the projection of the geodesic in $\mathcal{A}$.

4.5.9 Example: location-scale families

We will now provide a concrete example to demonstrate that this geometric view of PT can help design better annealing paths when the reference and target a nearly mutually singular.

Suppose $\mathcal{X} = \mathbb{R}$ and $p(x) = \exp(W(x)) \in \mathcal{P}(\mathcal{X})$, is normalized, with mean 0 and variance 1 and $W(x)$ is differentiable and symmetric. For $\eta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ we let

$$p_\eta(x) = \frac{1}{\sigma} p\left(\frac{x-\mu}{\sigma}\right).$$

Define $\mathcal{M} = \{p_\eta : \eta \in \Omega\}$, to be the location-scale family generated by p. This is a 2-parameter family where μ and σ are called the location and scale parameters corresponding to the mean and standard deviation of π_η. Notice that when $p = N(0,1)$ is the standard normal, then $p_\eta = N(\mu, \sigma^2)$, and $\mathcal{M}$ corresponds to the space of 1-dimensional Gaussians.

Suppose we wish to anneal between a reference that is a translation of the target, i.e. $\pi_0 = p_{\eta_0}$ and $\pi_1 = p_{\eta_1}$, where $\eta_0 = (\mu_0, \sigma)$ and $\eta_1 = (\mu_1, \sigma)$ respectively. When $z = |\mu_1 - \mu_0|/\sigma$ is large, this models the situation where the reference and target are nearly mutually singular. The simplest path between π_0 and π_1 is the translation path $\eta(\beta) = (\mu(\beta), \sigma)$, with $\mu(\beta) = (1-\beta)\mu_0 + \beta\mu_1$, that translates π_0 toward π_1 keeping the variance fixed in β. For example see Figure 4.1 (top). We remark that for general local-scale families, the translation path in $\mathcal{M}$ may not be equal to the linear path since $\pi_0^{1-\beta}\pi_1^\beta$ may not be in $\mathcal{M}$.

For location-scale families, we can explicitly compute the communication barrier for the translation path and the geodesic with respect to the SKL. Proposition 14 generalizes 9 from Section 4.1 and shows as $z \to \infty$, the global communication barrier for the translation path is $O(z)$, in contrast to the geodesic, which is $O(\log z)$. This implies that when the reference and target are nearly mutually singular, we can gain an exponential improvement in the round trip rate by choosing a different path.

Proposition 14. *Suppose the reference and target distributions are $\pi_0 = p_{\eta_0}$ and $\pi_1 = p_{\eta_1}$ corresponding to $\eta_0 = (\mu_0, \sigma)$ and $\eta_1 = (\mu_1, \sigma)$ in Ω, and define $z = |\mu_1 - \mu_0|/\sigma$. Then as $z \to \infty$,*

1. *the translation path has $\tau(\pi) = \Theta(1/z)$, and*
2. *the geodesic path has $\tau(\pi) = \Omega(1/\log z)$.*

For a fixed N, when z is sufficiently large, the intermediate distribution for the translation path $\{\pi_{\beta_n}\}_{n=0}^N$ are nearly mutually, leading to high rejection rates and hence low round trip rates (see Figure 4.1 (top)). In contrast the geodesic path moves mass from π_0 to π_1 while increasing variance for the intermediate distributions to increase their overlap and maintain a high rejection rate (see Figure 4.1 (bottom)). Although this is a toy model, by studying the geodesics of location-scale families, we will apply these insights to design path families robust to situations where the reference and target are nearly mutually singular in the following section.

Translation path

We can compute the global communication barrier explicitly.

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d} \log \pi_\beta(x)}{\mathrm{d}\beta} = -\frac{\mu_1 - \mu_0}{\sigma} \frac{\mathrm{d}W}{\mathrm{d}x} \left(\frac{x - \mu(\beta)}{\sigma} \right). \tag{4.34}$$

Since for all $\eta = (\mu, \sigma)$, we $X_\eta \sim p_\eta$, implies $(X_\eta - \mu)/\sigma \sim p$, we can substituting in (4.34) into (4.7), to get

$$\lambda(\beta) = \frac{|z|}{2} \mathbb{E}\left[\left| \frac{\mathrm{d}W}{\mathrm{d}x}(X) - \frac{\mathrm{d}W}{\mathrm{d}x}(X') \right| \right], \quad X, X' \sim p, \tag{4.35}$$

where $z = |\mu_1 - \mu_0|/\sigma$. Thus, the local communication is constant and the global communication barrier $\Lambda(\pi)$ equals (4.35). As $z \to \infty$, this implies $\Lambda(\pi) = \Theta(z)$, and hence the round trip rate $\tau(\pi) = \Theta(1/z)$.

Geodesic path

We will now find the geodesics and distance with respect to the SKL between π_0 and π_1 by using the analysis in Nielsen (2021, Appendix A). See also Costa et al. (2015, Section 2) for a similar computation for the Gaussian family. Using the Fisher information for the location-scale family $\mathcal{M}$ (Nielsen, 2020, Example 5), the Finsler/Riemannian metric for the SKL is,

$$G_{\mathrm{SKL}}(\eta) = \frac{1}{2} I(\eta) = \frac{1}{2\sigma^2} \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix},$$

where a, b are constants depending on p,

$$a^2 = \mathbb{E}_p\left[\left(\frac{\mathrm{d}W}{\mathrm{d}x}(X)\right)^2\right], \qquad b^2 = \mathbb{E}_p\left[\left(X\frac{\mathrm{d}W}{\mathrm{d}x}(X) + 1\right)^2\right].$$

For example, in the case where p is the standard Gaussian, we have $a = 1, b = 2$ (Costa et al., 2015). By re-scaling the location parameters by a/b, the Finsler metric for the rescaled parameters $\bar{\eta} = (\bar{\mu}, \bar{\sigma}) = (\frac{a}{b}\mu, \sigma)$ becomes,

$$G_{\mathrm{SKL}}(\bar{\eta}) = \frac{b^2}{2\bar{\sigma}^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \frac{b^2}{2}G_{\mathbb{H}}(\bar{\eta}), \tag{4.36}$$

where $G_{\mathbb{H}}(\bar{\eta})$ is the Riemannian metric for the Poincaré upper-half plane $\mathbb{H}$ [1]. If π is a path in $\mathcal{M}$ between π_0 and π_1 with location and scale parameters $\eta(\beta) = (\mu(\beta), \sigma(\beta))$, then (4.36) implies the length of π with respect to the SKL is,

$$\Lambda_{\mathrm{SKL}}(\pi) = \frac{b}{\sqrt{2}}\Lambda_{\mathbb{H}}(\bar{\eta}) \tag{4.37}$$

where $\Lambda_{\mathbb{H}}(\bar{\eta})$ is the length of the path of $\bar{\eta}(\beta) = (\frac{a}{b}\mu(\beta), \sigma(\beta))$ between $\bar{\eta}_0 = (\frac{a}{b}\mu_0, \sigma)$ and $\bar{\eta}_1 = (\frac{a}{b}\mu_1, \sigma)$ in $\mathbb{H}$. It is well known that the geodesics in $\mathbb{H}$ are vertical half-lines or half-circles centered at $\bar{\sigma} = 0$ (Lee, 2006, Chapter 3), so the geodesics in $\mathcal{M}$ correspond to vertical half-lines and half-ellipses in Ω with eccentricity a/b. By taking the infimum over annealing paths π in (4.37), the distance with respect to the SKL can be written term of the distance $\bar{\eta}_0$ and $\bar{\eta}_1$ in $\mathbb{H}$,

$$\begin{aligned} d_{\mathrm{SKL}}(\pi_0, \pi_1) &= \frac{b}{\sqrt{2}}d_{\mathbb{H}}(\bar{\eta}_0, \bar{\eta}_1) \\ &= b\sqrt{2}\log\left(\frac{a}{b}z + \sqrt{\frac{a^2}{b^2}z^2 + 4}\right). \end{aligned}$$

[1] The Poincaré upper-half plane, $\mathbb{H}$, is the subset $\mathbb{H} = \mathbb{R} \times \mathbb{R}_+ \subset \mathbb{R}^2$ equipped with the Riemannian metric

$$G_{\mathbb{H}}(\bar{\eta}) = \frac{1}{\bar{\sigma}^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Given $\bar{\eta}_0, \bar{\eta}_1 \in \mathbb{H}$, the distance in $\mathbb{H}$ equals,

$$d_{\mathbb{H}}(\bar{\eta}_0, \bar{\eta}_1) = 2\log\left(\frac{\sqrt{(\bar{\mu}_1 - \bar{\mu}_0)^2 + (\bar{\sigma}_1 - \bar{\sigma}_0)^2} + \sqrt{(\bar{\mu}_1 - \bar{\mu}_0)^2 + (\bar{\sigma}_1 + \bar{\sigma}_0)^2}}{2\sqrt{\bar{\sigma}_0\bar{\sigma}_1}}\right).$$

Since $d_{r^2}(\pi_0, \pi_1) \leq d_{\mathrm{SKL}}(\pi_0, \pi_1)$, as $z \to \infty$, we have the distance $d_{r^2}(\pi_0, \pi_1) = O(\log(z))$, and hence the round trip rate of the geodesic path is $\tau(\pi) = \Omega(1/\log z)$.

4.6 Spline annealing path family

In this section, we develop an annealing path family $\mathcal{A} = \{\pi^\phi : \phi \in \Phi\}$ that offers a practical and flexible improvement upon the traditionally used linear path. Given a fixed reference $\pi_0 \propto \exp(W_0)$ and target $\pi_1 \propto \exp(W_1)$ distributions, we begin with the practical desiderata for an annealing path family to be useful in the context of parallel tempering. First, the traditional linear path, $\pi_\beta \propto \exp(W_\beta)$ where $W_\beta = (1-\beta)W_0 + \beta W_1$, should be a member of the family, so that one can achieve at least the round trip rate provided by that path. Second, the family should be broadly applicable and not depend on precise details of either π_0 or π_1 or the state space $\mathcal{X}$. Finally, using the Gaussian example from Figure 4.1 and Section 4.5.9 as insight, the family should enable the path to smoothly vary from π_0 to π_1 while inflating/deflating the variance as necessary.

Since we want our annealing path family to include the linear path, we begin our exploration by first studying the structure of the linear path. In particular, for the linear path the log-likelihood W_β is a convex combination of the form $\eta_0(\beta)W_0 + \eta_1(\beta)W_1$ where $\eta_0(\beta) = 1-\beta$ and $\eta_1(\beta) = \beta$. We will relax the constraint forcing the weights $\eta_0(\beta)$ and $\eta_1(\beta)$ to sum to 1 since our desiderata do not require this. This is the motivation for the construction of the exponential annealing family.

4.6.1 Exponential annealing family

Suppose we have a fixed reference $\pi_0(x) \propto \exp(W_0(x))$ and target $\pi_1(x) \propto \exp(W_1(x))$. For $\eta = (\eta_0, \eta_1) \in \mathbb{R}^2$, and $W(x) = (W_0(x), W_1(x))$, define $p_\eta \in \mathcal{P}(\mathcal{X})$, such that,

$$p_\eta(x) = \frac{1}{Z(\eta)} \exp(\eta^T W(x)),$$

where $Z(\eta) = \int_{\mathcal{X}} \exp(\eta^T W(x)) \mathrm{d}x$ when defined. Define $\Omega = \{\eta \in \mathbb{R}^2 : Z(\eta) < \infty\}$, then $\mathcal{M} = \{p_\eta : \eta \in \Omega\}$ is an exponential family with sufficient statistics $W(x)$ and natural parameters η. Intuitively, η_0 and η_1 represent the level of annealing for the reference and target, respectively. By Proposition 15, under some mild integrability conditions, $\mathcal{M}$ is an annealing family making it amenable to

parallel tempering so we shall refer to $\mathcal{M}$ as the *exponential annealing family*. In particular linear paths between η_0 and η_1 in Ω correspond to linear paths between π_{η_0} and π_{η_1} in $\mathcal{M}$.

Proposition 15. *If $\mathcal{M}$ is the exponential familiy generated by π_0 and π_1 and if for all compact $K \subset \Omega$,*

$$\sup_{\eta \in K} \mathbb{E}_\eta[\|W\|^3] < \infty,$$

then $\mathcal{M}$ is an annealing family.

See Appendix A.2.4 for the proof.

4.6.2 Spline annealing path family

Let $\mathcal{A}$ be the set of regular annealing paths for $\mathcal{M}$. For each $\pi \in \mathcal{A}$, there is a piece-wise twice continuously differentiable $\eta : [0,1] \to \Omega$ such that $\pi_\beta(x) \propto \exp(\eta(\beta)^T W(x))$. When the annealing schedule $\mathcal{B}_N$ is fixed, the analysis in Section 4.3 shows that PT cannot distinguish between π and the spline approximation $\pi^N_\beta \propto \exp(W^N_\beta) \in \mathcal{A}$ defined by (A.38). Since linear paths between p_η and $p_{\eta'} \in \mathcal{M}$ correspond to the linear path connecting η and η' in Ω, we have $W^N_\beta(x) = \eta^N(\beta)^T W(x)$ where η^N a linear spline in Ω with N knots. This motivates our construction of the spline annealing path family.

Suppose $\phi = (\eta_0, \ldots, \eta_K) \in \Omega^{K+1} = \Phi_K$, such that $\eta_0 = (1,0)$ and $\eta_K = (0,1)$. Then, define the K-knot linear spline $\eta^\phi : [0,1] \to \Omega$ that takes on values η_k at $\beta = k/K$ and linearly interpolates between η_{k-1} and η_k for $\beta \in [\frac{k-1}{K}, \frac{k}{K}]$. More precisely,

$$\eta^\phi_\beta = (k - K\beta)\eta_{k-1} + (K\beta - k + 1)\eta_k, \quad \beta \in \left[\frac{k-1}{K}, \frac{k}{K}\right].$$

We will denote the corresponding annealing path π^ϕ_β in $\mathcal{M}$ with log-likelihood $W^\phi_\beta(x) = \eta^\phi(\beta)^T W(x)$. The K-knot *spline annealing path family* is defined as the set of K-knot linear spline paths denoted $\mathcal{A}_K = \{\pi^\phi\}_{\phi \in \Phi_K}$.

Validity: If $p_{\eta_0}, p_{\eta_1} \in \mathcal{M}$, then so is the linear path between them. In particular we always have the linear path is a member of $\mathcal{A}_K$ for all K, ensuring we do no worse than the PT without path

tuning.

Convexity: Furthermore, the convexity of Ω implies tuning the knots $\phi = (\eta_0, \ldots, \eta_K) \in \Phi_K$ involves optimization within a convex constraint set. In practice, we enforce also that the knots $\eta_k = (\eta_{k,0}, \eta_{k,1})$ are monotone in each component—i.e., the first component monotonically decreases, $1 = \eta_{0,0} \geq \eta_{1,0} \geq \cdots \geq \eta_{K,0} = 0$ and the second increases, $0 = \eta_{0,1} \leq \eta_{1,1} \leq \cdots \leq \eta_{K,1} = 1$ —such that the path of distributions always moves from the reference to the target. Because monotonicity constraint sets are linear and hence convex, the overall monotonicity-constrained optimization problem has a convex domain.

Flexibility: Suppose $\pi_\beta \propto \exp(\eta(\beta)^T W) \in \mathcal{A}$. Given a large enough number of knots K, the spline annealing family well-approximates π_β. By Taylor's theorem,

$$\inf_{\phi \in \Phi_K} \|\eta^\phi(\beta) - \eta(\beta)\|_\infty \leq \frac{1}{4K^2} \left\| \frac{d^2\eta}{d\beta^2} \right\|_\infty . \tag{4.38}$$

Let $\pi^K_\beta \propto \exp(\eta^K(\beta)^T W) \in \mathcal{A}_K$ be the path that attains the infimum in (4.38) and corresponding to the projection of π onto $\mathcal{A}_K$. When $K = 1$, we note that $\pi^K_\beta = \pi_0^{1-\beta}\pi_1^\beta$ always reduces to the linear path and $K \to \infty$ we have π^K_β converges to π_β in total variation. Therefore as K increases, π^K interpolates between the linear path and a fully general path $\pi \in \mathcal{A}$. In particular, PT cannot distinguish between path $\pi \in \mathcal{A}$ and the linear spline path with knots $\pi_{\beta_0}, \ldots, \pi_{\beta_N}$. Therefore, in PT, we can non-asymptotically reconstruct any path using at most $K = N$ knots.

4.6.3 Tuning K

As we established above, as K increases, the flexibility of the annealing path family $\mathcal{A}_K$ increases up to $K = N$. Since $\eta_0 = (1, 0), \eta_1 = (0, 1)$ are fixed, $\mathcal{A}_K$ can be parameterized by optimizing the SKL objective over knots $\eta_1, \ldots, \eta_{K-1}$, with $2(K-1)$ free parameters. In particular, this suggests we do not want to choose $K = O(N)$ since the optimization problem's complexity would grow linearly with the number of chains, which would not be scalable when N is large. We want to pick K large enough so that $\mathcal{A}_K$ is reasonably flexible but small enough so that the optimization problem is computationally tractable. In practice, the most significant improvements are found from $K = 1$ to $K = 2$, corresponding to a single free knot, and after this, the gains are marginal. The optimized

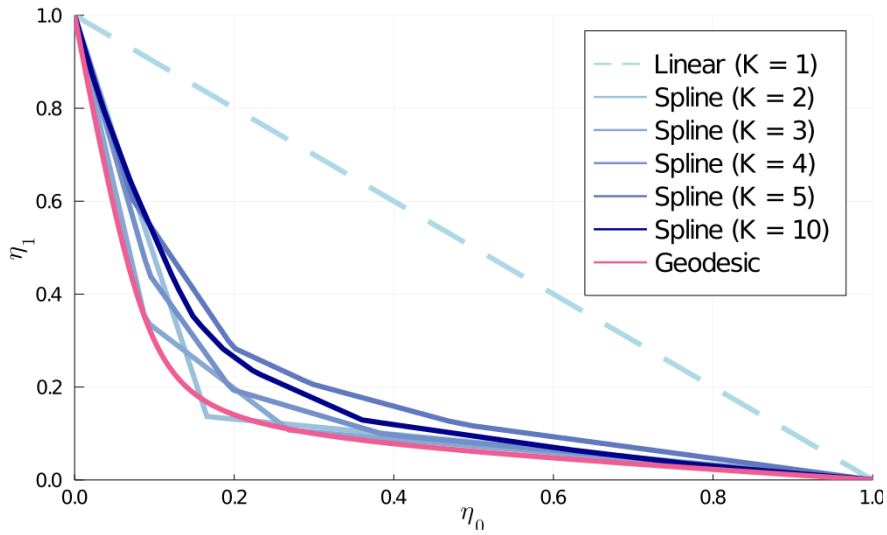

Figure 4.2: The spline path for $K = 1, 2, 3, 4, 5, 10$ knots in the exponential annealing family generated by $\pi_0 = N(-1, 0.5)$ and $\pi_1 = N(1, 0.5)$. The smooth pink line is the geodesics with respect to the SKL.

path becomes smoother for $K > 2$, but the performance improvements become marginal. We want to pick $K \ll N$. As an example, we see in Figure 4.2, as K increases, the optimized spline path approximates the geodesic for the exponential annealing family generated by the Gaussians. See Syed et al. (2021b) for tuning guidelines for choosing K.

4.6.4 Example: Gaussian

The exponential annealing family generated by the Gaussians is also a location-scale family, where we have explicit characterization of the linear path and geodesics ($a = 1, b = \sqrt{2}$) from Section 4.5.9. We can use this to show to validate the theory and performance of Algorithm 6. Figure 4.2 illustrates the behaviour of optimized spline paths for a Gaussian reference and target. The path takes a convex curved shape, starting at the bottom right point of the Figure 4.2. This path corresponds to increasing the variance of the reference, shifting the mean from reference to target, and finally decreasing the variance to match the target. With more knots, this process happens more smoothly.

When $\pi_0 = N(\mu_0, \sigma^2)$ and the target $\pi_1 = N(\mu_1, \sigma^2)$, we can use (4.35) to compute the

communication barrier for the linear path analytically,

$$\Lambda(\pi) = \frac{z}{2}\mathbb{E}[|Z - Z'|], \quad Z, Z' \sim N(0, 1),$$

where $z = |\mu_1 - \mu_0|/\sigma$. Since $Z - Z' \sim N(0, 2)$, we have $\mathbb{E}[|Z - Z'|]$ is the mean of a folded-normal distribution and equal 1.128.... This implies that the linear path for the Gaussian has communication barrier $\Lambda \approx .564z$.

Suppose $\pi_0 = N(-1, 0.01^2)$ and $\pi_1 = N(1, 0.1^2)$, with $z = 200$. We used Algorithm 6 with $N = 50$ parallel chains initialized at the linear path between π_0 and π_1 with the uniform schedule. The results of these experiments are shown in Figure 4.3. Figure 4.3 (left) shows the number of round trips as a function of the number of scans. Note that the global communication barrier $\Lambda \approx 113$ for the linear path is much larger than N. Algorithms based on linear paths incurred rejection rates of nearly one for most chains, resulting in no round trips.

One can see that PT using the spline annealing family outperforms PT using the linear annealing path across all numbers of knots tested. Moreover, the slope of these curves demonstrates that PT with the spline annealing family exceeds the theoretical upper bound of round trip rate for the linear path. The largest gain is obtained from $K = 1$ (linear) to $K = 2$. For all the examples, increasing the number of knots to more than $K > 2$ leads to marginal improvements. Figure 4.3 (right) shows the value of the surrogate SKL objective and non-asymptotic communication barrier. In particular, this demonstrates that the SKL provides a surrogate objective that is a reasonable proxy for the non-asymptotic communication barrier.

See Section 5 of Syed et al. (2021b) for a more thorough exploration into the practical implementation of Algorithm 6, and path tuning more generally. In particular, Syed et al. (2021b) empirically demonstrates that Algorithm 6 can be applied as broadly as non-reversible PT with the linear path but can also outperform the theoretically optimal performance of the linear path potentially by large margins. Moreover, they show PT with path tuning is also more robust for hard problems where π_0 and π_1 are nearly mutually singular and when the model is high dimensional.

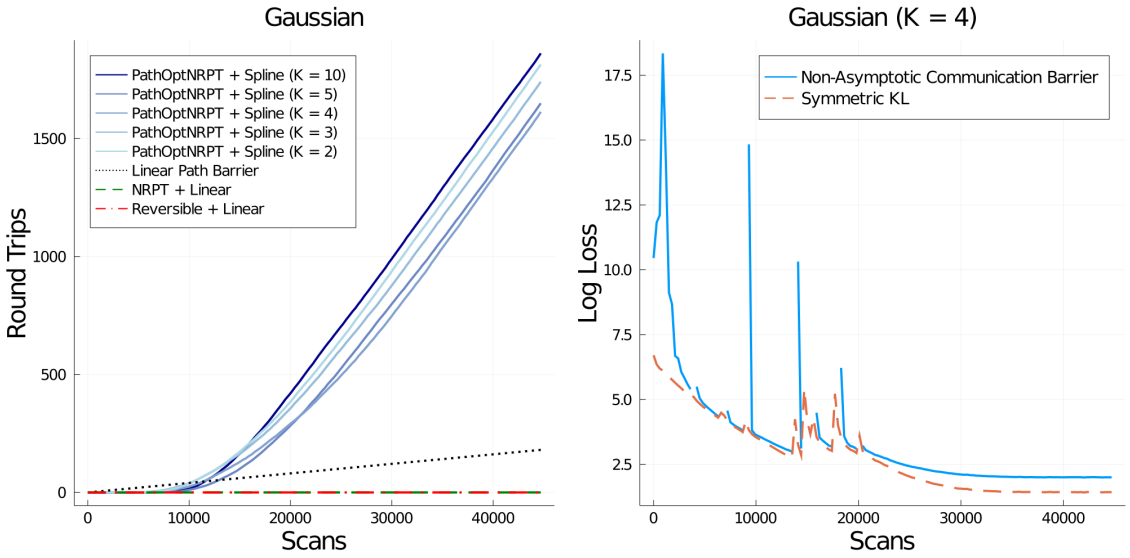

Figure 4.3: (Left) Cumulative round trips averaged over 10 runs for the spline path with $K = 2, 3, 4, 5, 10$ (solid blue), non-reversible PT using a linear path (dashed green), and reversible PT with linear path (dash/dot red). The slope of the lines represent the round trip rate. We observe large gains going from linear to non-linear paths ($K > 1$). For all values of $K > 1$, the optimized spline path substantially improves on the theoretical upper bound on round trip rate possible using linear path (dotted black). (Right) Non-asymptotic communication barrier $\Lambda(\pi^\phi, \mathcal{B}_N)$ (solid blue) and symmetric KL objective $\text{SKL}(\pi^\phi, \mathcal{B}_N)$ (dash orange) in log-scale as a function of iteration for one run of PathOptimizedNRPT + Spline ($K = 4$ knots).

Chapter 5

Scaling limit for parallel tempering

If you just focus on the smallest details, you never get the big picture right.

— Leroy Hood

5.1 Introduction

Recall the difference between non-reversible and reversible PT is just the difference in the communication kernel. The difference between DEO and SEO communication that defines the two variants of PT is simply the difference between alternating between even and odd swaps or choosing them at random. We have seen that DEO dominates its SEO in terms of round trip rate for any choice of N and schedule (Corollary 2), and the performance gap widens as N increases (Theorem 6). In particular, the round trip rate in the reversible case deteriorates to zero, in contrast to DEO, which improves in performance as N increases. So far, we have focused on leveraging these quantitative differences in Chapters 3 and 4 to develop guidelines and methodology to tune DEO. It remains to be addressed why such seemingly trivial algorithmic change can lead to such dramatic change in both non-asymptotic and asymptotic performance.

The DEO scheme first introduced in Okabe et al. (2001) was presumably devised on algorithmic grounds (it performs the maximum number of swap attempts in parallel) and provided no theoretical justification or reference to non-reversibility. Lingenheil et al. (2009) first identified a qualitative difference between the two communication schemes and empirically showed that DEO required different tuning guidelines than SEO. The arguments given in Lingenheil et al. (2009) to explain the superiority of DEO communication over various PT algorithms rely on a misleading assumption, namely a diffusive scaling limit for the index process. Figures 1.4 and 2.3 suggest that the index process behaves qualitatively differently as N increases for reversible and non-reversible PT. DEO does not exhibit the same diffusive behaviour for large N that SEO does. In particular, Figure 2.3

suggests that when time is scaled by a factor of N, the law of the index process stabilizes. This section aims to investigate these differences by identifying the scaling limits of the index process as N increases. Such limits exist under assumptions (A1') and (A2') specified in Section 4.2.4.

Scaling limits are a powerful tool to understand the behaviour of the structure of a family of probabilistic objects. Often we want to say that our random object is similar to a simpler object that you can describe precisely. By appropriately taking limits of a scaled object, we can provide approximate descriptions of what it should look like when you look from a "large distance" or "zoom-out". Scaling limits are widely used in the literature to study the qualitative behaviour of Monte Carlo algorithms in various asymptotic regimes (Gelman et al., 1997; Roberts and Rosenthal, 2001; Beskos et al., 2013; Bierkens and Roberts, 2017; Deligiannidis et al., 2018). Previous theoretical studies for PT analyzed the high-dimensional scaling behaviour of the acceptance probability based on a target consisting of a product of independent components of increasing dimension (Atchadé et al., 2011), or an increased swap frequency between exploration steps using the Langevin exploration kernel (Dupuis et al., 2012). We will formally identify the scaling behaviour of the index process as N increases encoding the limiting behaviour of PT as parallel computation resources increase.

5.2 Scaling limits of the index process

We suppose π is a regular annealing path defined in Section 4.3.1, with schedule generated by γ. Suppose (I_t, ϵ_t) is the index process for an annealing schedule $\mathcal{B}_N$ generated by γ. To establish a scaling limit for (I_t, ϵ_t), it will be convenient to work in a continuous time setting. To do this, we suppose the times that PT scans occur are distributed according to a Poisson process $\{M(\cdot)\}$ with mean μ_N. The number $M(t)$ of PT iterations that occur by time $t \geq 0$ thus satisfies $M(t) \sim \text{Poisson}(\mu_N t)$. We define the *scaled index process* by $Z^N(t) = (W^N(t), \epsilon^N(t))$ where $W^N(t) := I_{M(t)}/N$ is the *scaled index* and $\epsilon^N(t) := \epsilon_{M(t)}$ is the *scaled lifting parameter*. This corresponds to "speeding" up the index process by a factor of μ_N.

Define the piecewise-deterministic Markov process (PDMP) (Davis, 1993) $Z(t) = (W(t), \epsilon(t))$ on $[0, 1] \times \{-1, 1\}$ as follows: $W(t)$ moves in $[0, 1]$ with velocity $\epsilon(t)$ and the sign of $\epsilon(t)$ is reversed at an inhomogeneous rate $\lambda(\gamma(W(t))\dot{\gamma}(W(t))$ or when $W(t)$ hits a boundary; see Bierkens et al. (2018) for a discussion of PDMP on restricted domains. The process Z corresponds to an inhomogeneous

persistent random walk with reflective boundary conditions; see Masoliver et al. (1992) for a description of the Fokker-Plank equation and first passage times.

Theorem 16. *Suppose π is a regular annealing path, and $\mathcal{B}_N$ is the family of schedules generated by γ, and assumptions (A1') and (A2') hold, then*

(a) For reversible PT if $\mu_N = N^2$ and if $W^N(0)$ converges weakly to $W(0)$ then W^N converges weakly to a diffusion W, where W is a Brownian motion on $[0,1]$ with reflective boundary conditions. The process W admits $\mathrm{Unif}([0,1])$ as stationary distribution.

(b) For non-reversible PT if $\mu_N = N$ and if $Z^N(0)$ converges weakly to $Z(0)$, then Z^N converges weakly to the PDMP Z with initial condition $Z(0)$. The process Z admits $\mathrm{Unif}([0,1] \times \{-1,1\})$ as stationary distribution.

Notation: We would like to emphasize, for the remainder of this chapter $Z(t)$ and $W(t)$ will denote the PDMP we just described and a Brownian motion reflected in the boundaries of $[0,1]$ respectively rather than log-likelihood, and normalizing constant.

Reversible PT

Theorem 16 implies for reversible PT that if we speed time by a factor of N^2, then the index process scales to a diffusion W (see Figure 5.1). For large N, the diffusivity of the index process dominates the reduction in the rejection rate. Since the scaling limit W is independent of the annealing path and schedule, even a perfectly tuned reversible PT algorithm performance collapses when N is large. As discussed in Section 3.3.3, for a path with communication barrier Λ, the number of chains required to maintain a constant rejection rate is $\Theta(\Lambda)$ as $\Lambda \to \infty$. It follows for challenging problems when Λ is large, SEO communication is particularly fragile since the diffusivity intrinsic to reversible PT will make it exceedingly difficult for round trips to occur.

Non-reversible PT

By speeding up non-reversible PT by a factor of N, the scaled index process converges to the PDMP Z. The limit depends on the path through the communication barrier and the schedule through γ. This scaling behaviour explains why DEO communication exhibits such drastically

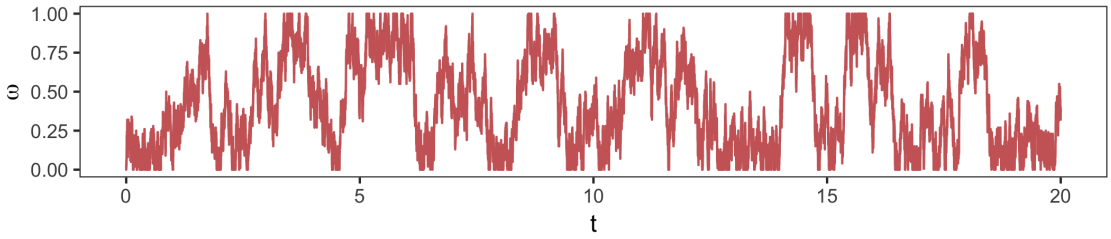

Figure 5.1: Sample trajectory of reversible scaling limit $W(t)$ corresponding to Brownian motion on $[0, 1]$ with reflective boundary condition.

different performance and tuning guidelines than its reversible counterpart. See Figure 5.2 for sample trajectories of Z for different values of Λ.

In particular, in light of Theorem 16, we have a new interpretation of the communication barrier: the local communication barrier λ governs the instantaneous rate of reflection of $W(t)$, and since for any generator γ,

$$\int_0^1 \lambda(\gamma(w))\dot{\gamma}(w)\mathrm{d}w = \Lambda,$$

the global communication barrier Λ is the total rate of reflection. Recall the optimal schedule derived in Chapters 3 and 4 is generated by $\gamma(w) = F^{-1}(w)$ for $F(\beta) = \int_0^\beta \lambda(u)\mathrm{d}u/\Lambda$. The optimal schedule generator ensures the rate of reflection is constant, i.e. $\lambda(\gamma(w))\dot{\gamma}(w) = \Lambda$ for all $w \in [0, 1]$ and $\epsilon(t)$ changes sign at a constant rate Λ.

5.3 Scaled index process

For convenience, we will use $z = (w, \epsilon) \in [0, 1] \times \{-1, 1\}$ to be a *scaled index*. Define $C(\mathbb{R}_+, \mathcal{S})$ and $D(\mathbb{R}_+, \mathcal{S})$ to be sets of functions $f : \mathbb{R}_+ \to \mathcal{S}$ that are continuous and càdlàg respectively.

The process $Z^N \in D(\mathbb{R}_+, [0, 1]\times\{-1, 1\})$ takes values on the discrete set $\{0, 1/N, \ldots, 1\}\times\{-1, 1\}$ and is only well-defined when $Z^N(0) = z_0 \in \{0, 1/N, \ldots, 1\} \times \{-1, 1\}$. 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 let $T_1, T_2, \ldots$ be the iteration times generated by the Poisson process $M(t)$. We construct $Z^N(t)$ as follows: define $Z^N(t) = z_i$ for $t \in [T_i, T_{i+1})$ and update $z_{i+1}|z_i$ via a transition kernel dependent on the communication scheme. We determine this transition kernel

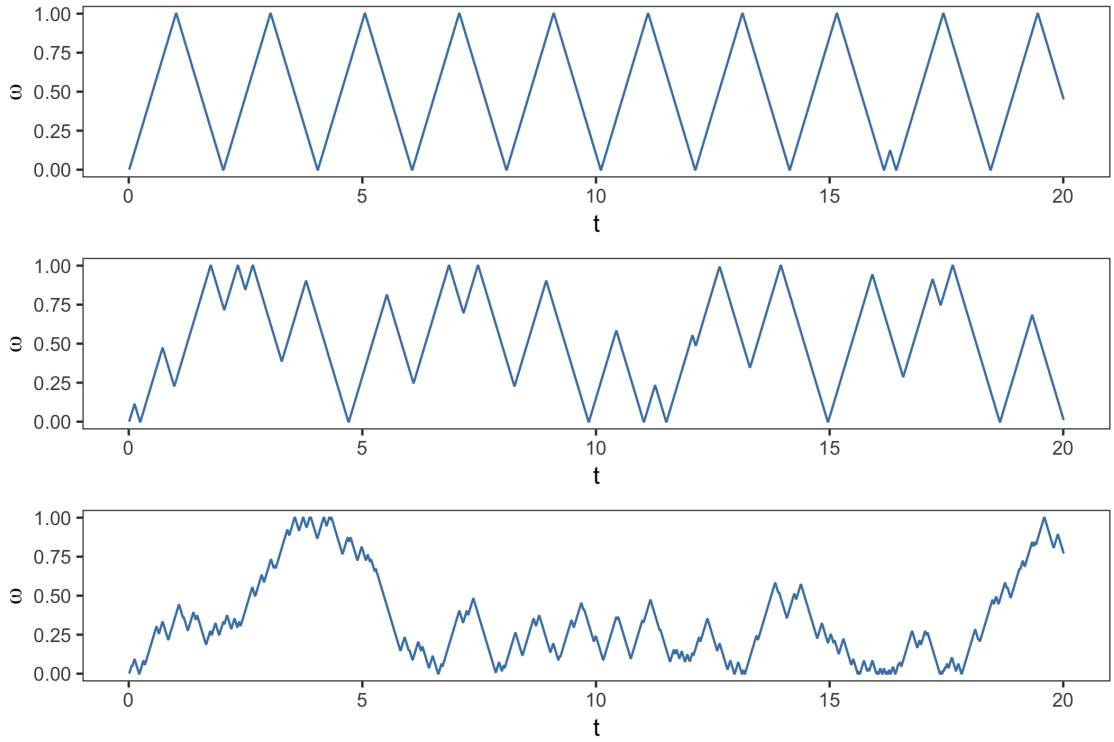

Figure 5.2: Sample trajectories of $W(t)$ where $Z(t) = (W(t), \epsilon(t))$ under an optimal schedule generated by $\gamma = F^{-1}$ for $F(\beta) = \Lambda(\beta)/\Lambda$ for $\Lambda = 0.1$ (top), $\Lambda = 1$ (middle), and $\Lambda = 10$ (bottom) respectively.

mirroring the construction from Section 3.1.4.

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

$$\Phi_-^N(w) = \begin{cases} w - \frac{1}{N} & w \in \left[\frac{1}{N}, 1\right], \\ \frac{1}{N} - w & w \in \left[0, \frac{1}{N}\right), \end{cases} \tag{5.1}$$

and similarly,

$$\Phi_+^N(w) = \begin{cases} w + \frac{1}{N} & w \in \left[0, 1 - \frac{1}{N}\right], \\ 1 - \left(\frac{1}{N} - (1 - w)\right) & w \in \left(1 - \frac{1}{N}, 1\right]. \end{cases} \tag{5.2}$$

Intuitively $\Phi_\epsilon^N(w)$ 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.

5.3.1 Scaled index process for reversible PT

Under the SEO communication scheme, if $z_i = (w_i, \epsilon_i) \in \{0, \frac{1}{N}, \dots, 1\} \times \{-1, 1\}$, then we have $w_{i+1} = \Phi^N_{\epsilon_i}(w_i)$ if a swap successfully occurred and $w_{i+1} = w_i$ otherwise. In both cases, $\epsilon_{i+1} \sim \text{Unif}\{-1, +1\}$. Since $\Phi^N_\epsilon(w)$ is not only well-defined for $w \in \{0, \frac{1}{N}, \dots, 1\}$ but for $w \in [0, 1]$, we naturally extend this construction to any $w \in [0, 1]$.

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

$$w_{i+1}|w_i, \epsilon_i \sim \begin{cases} \Phi^N_{\epsilon_i}(w_i) & \text{with probability } s\left(\gamma(w_i), \gamma(\Phi^N_{\epsilon_i}(w_i))\right), \\ w_i & \text{otherwise.} \end{cases}$$

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

From Theorem 19.2 in Kallenberg (2002), the infinitesimal generator for W^N with SEO communication is

$$\mathcal{L}_{W^N} f(w) = \frac{\mu_N}{2} \sum_{\epsilon \in \{\pm 1\}} \left(f(\Phi^N_\epsilon(w)) - f(w)\right) s\left(\gamma(w), \gamma(\Phi^N_\epsilon(w))\right), \tag{5.3}$$

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 $\mathcal{D}(\mathcal{L}_{W^N}) = C([0, 1])$.

5.3.2 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 $z = (w, \epsilon)$,

$$\Phi^N(z) = \begin{cases} (\Phi^N_\epsilon(w), \epsilon) & \text{if } \Phi^N_\epsilon(w) = w + \frac{\epsilon}{N}, \\ (\Phi^N_\epsilon(w), -\epsilon) & \text{otherwise,} \end{cases}$$

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

$$R(z) = (w, -\epsilon).$$

Under the DEO scheme, if $z_n = (w_n, \epsilon_n) \in \{0, \frac{1}{N}, \ldots, 1\} \times \{-1,1\}$, then we have $z_{n+1} = \Phi^N(z_n)$ when a swap is accepted 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_{i+1} according to the transition kernel,

$$z_{i+1}|z_i \sim \begin{cases} \Phi^N(z_i) & \text{with probability } s\left(\gamma(w_i), \gamma(\Phi^N_{\epsilon_i}(w_i))\right), \\ R(z_i) & \text{otherwise.} \end{cases}$$

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

Analogously to the reversible case, under DEO communication, the infinitesimal generator for Z^N is

$$\mathcal{L}_{Z^N} f(z) = \mu_N \left(f(\Phi^N(z)) - f(z)\right) s\left(\gamma(w), \gamma(\Phi^N_\epsilon(w))\right) + \mu_N \left(f(R(z)) - f(z)\right) r\left(\gamma(w), \gamma(\Phi^N_\epsilon(w))\right),$$

where $z = (w, \epsilon)$ 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 $w_0 \in \{0,1\}$.

5.4 Proof of scaling limit for reversible PT

We will prove Theorem 16(a) by using Theorem 17.25 from Kallenberg (2002).

Theorem 17 (Trotter, Sova, Kurtz, Mackevičius)**.** *Let $X, X^1, X^2, \ldots$ be Feller processes defined on a state space S with generators $\mathcal{L}, \mathcal{L}_1, \mathcal{L}_2, \ldots$ respectively. If D is a core for $\mathcal{L}$, then the following statements are equivalent:*

1. *If $f \in D$, there exists $f_N \in \mathcal{D}(\mathcal{L}_N)$ 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 17 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

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

and $\mathcal{L}_N = \mathcal{L}_{W^N}$ defined in (5.3), which we recall here for the reader's sake

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

with $\Phi_\pm^N(w)$ defined in (5.1), (5.2) and taking $\mu_N = N^2$. Recall from the discussion just before (5.3) that $\mathcal{L}_{W^N}$ defines a Feller semigroup.

First notice that in Kallenberg (2002), 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 if 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.

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

$$\mathcal{L}_W f(w) = \frac{1}{2} \frac{\mathrm{d}^2 f}{\mathrm{d}w^2}, \tag{5.4}$$

where the domain $\mathcal{D}(\mathcal{L}_W)$ is the set of functions $f \in C^2([0,1])$ 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.

Proposition 18. *W is a Feller process with generator $\mathcal{L}_W$ defined by* (5.4).

See Appendix A.3.1 for the proof.

Now we can apply Theorem 17 to prove Theorem 16(a). We only need to check the first condition of Theorem 17. In this direction, first note that by definition $\Phi_\pm^N(w) = w \pm 1/N$ for $w \in [1/N, 1 - 1/N]$. Thus in this case using a Taylor expansion we have for $w_-^* \in [w - 1/N, w]$ and

$w_+^* \in [w, w+1/N]$ that for any $f \in \mathcal{D}(\mathcal{L}_W)$,

$$\begin{aligned} f(\Phi_+^N(w)) - 2f(w) + f(\Phi_-^N(w)) &= f(w) + \frac{1}{N}f'(w) + \frac{1}{2N^2}f''(w_+^*) \\ &\quad + f(w) - \frac{1}{N}f'(w) + \frac{1}{2N^2}f''(w_-^*) - 2f(w) \\ &= \frac{1}{2N^2}\left(f''(w_+^*) + f''(w_-^*)\right). \end{aligned}$$

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

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

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

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

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 case being similar. Let $w \in [0, 1/N)$ in which case, since $f'(0) = 0$, for $w^*, w_-^*, w_+^* \in [0, 2/N]$

$$\begin{aligned} f(\Phi_+^N(w)) - 2f(w) + f(\Phi_-^N(w)) &= f(0) + \Phi_+^N(w)f'(0) + \frac{1}{2}\left[\Phi_+^N(w)\right]^2 f''(w_+^*) \\ &\quad + f(0) + \Phi_-^N(w)f'(0) + \frac{1}{2}\left[\Phi_-^N(w)\right]^2 f''(w_-^*) \\ &\quad - 2f(0) - 2f'(0)w - 2\frac{f''(w^*)}{2}w^2 \\ &= \frac{f''(0)}{2}\left\{\left[\Phi_+^N(w)\right]^2 + \left[\Phi_-^N(w)\right]^2 - 2w^2\right\} + o\left(N^{-2}\right), \end{aligned}$$

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

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

Using Theorem 10 we see that for some constant $C > 0$,

$$\sup_{w\in[0,1]} \left| s\left(\gamma(w), \gamma(\Phi_{\pm}^N(w))\right) - 1 \right| \leq C \sup_w \left|\gamma(w) - \gamma\left(\Phi_{\pm}^N(w)\right)\right| \leq \frac{C\|\dot{\gamma}\|_\infty}{N},$$

and therefore

$$\begin{aligned}
\mathcal{L}_N f(w) &= \frac{N^2}{2} \sum_{\epsilon\in\{\pm 1\}} \left(f(\Phi_\epsilon^N(w)) - f(w)\right) s\left(\gamma(w), \gamma(\Phi_\epsilon^N(w))\right) \\
&= \frac{N^2}{2} \sum_{\epsilon\in\{\pm 1\}} \left(f(\Phi_\epsilon^N(w)) - f(w)\right) \left[1 + o(N^{-1})\right] \\
&= \frac{N^2}{2} \left(f(\Phi_+^N(w)) - 2f(w) + f(\Phi_-^N(w))\right) \left[1 + o(N^{-1})\right] \\
&= \frac{N^2}{2} \frac{f''(w)}{N^2} \left[1 + o(1)\right],
\end{aligned}$$

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

5.5 Proof of scaling limit for non-reversible PT

Define $Z \in C(\mathbb{R}_+, [0,1] \times \{-1,1\})$ to be the PDMP on $[0,1] \times \{-1,1\}$ given by $Z(t) = (W(t), \epsilon(t))$ where $W(t)$ moves in $[0,1]$ with velocity $\epsilon(t)$ and the sign of $\epsilon(t)$ is reversed at the arrivals times of a non-homogeneous Poisson process of rate $\lambda(\gamma(W(t)))\dot{\gamma}(W(t))$ or when $W(t)$ reaches the boundary $\{(0,-1),(1,+1)\}$. The infinitesimal generator of Z is given by

$$\mathcal{L}_Z f(z) = \epsilon \frac{\partial f}{\partial w}(z) + \lambda(\gamma(w))\dot{\gamma}(w) \left(f(R(z)) - f(z)\right),$$

for any $f \in \mathcal{D}(\mathcal{L}_Z)$, 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 $w_0 \in \{0,1\}$.

We will prove Theorem 16(b) in a slightly round about way. We will define the 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

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

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

$$Q^N(\theta, \mathrm{d}\theta') = s\left(\tilde{\gamma}(\theta), \tilde{\gamma}(\Sigma^N(\theta))\right) \delta_{\Sigma^N(\theta)}(\mathrm{d}\theta') + r\left(\tilde{\gamma}(\theta), \tilde{\gamma}(\Sigma^N(\theta))\right) \delta_{2\pi-\theta}(\mathrm{d}\theta'),$$

where

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

Define the map

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

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

$$\tilde{\lambda}(\theta) = (\lambda \circ \gamma)(\phi^1(\theta))\dot{\gamma}(\phi^1(\theta)),$$

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

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

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 $\{U(\cdot)\}$.

Proposition 19. *The process U defined above is a Feller process, its infinitesimal generator is given by*

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

with domain

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

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

See Appendix A.3.2 for the proof.

Proposition 20. *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 17. The generator of U_N is given by

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

We will consider the two terms separately. To this end notice that by Theorem 10, the boundedness of λ and the fact that $\gamma \in C^1[0, 1]$,

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

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

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

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 $N \to \infty$

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

Next, using Theorem 10 we have that

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

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

$$N\left[f(-\theta) - f(\theta)\right] r\left(\tilde{\gamma}(\theta), \tilde{\gamma}(\Sigma^N(\theta))\right) \to \tilde{\lambda}(\theta)\left[Qf(\theta) - f(\theta)\right],$$

uniformly in θ. □

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 20 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 $\{0, \pi, 2\pi\}$, we will be using (Billingsley, 2013, 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 $k \in \mathbb{N}$ and $0 < t_1 < \cdots < t_k$ we want

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

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 $\$^1$.

Relative compactness of $\{Z^N(\cdot)\}_N$ can be easily seen to follow from the compact containment condition (Ethier and Kurtz, 2009, Remark 3.7.3). This combined with convergence of the finite dimensional distributions of Z^N to those of Z concludes the proof.

Chapter 6

Conclusions

A good tool improves the way you work. A great tool improves the way you think.

— Jeff Duntemann

6.1 Summary of contributions

We formally introduced the PT algorithm in Chapter 2 and distinguish between PT with reversible and non-reversible communication. We then characterized their differences through the dynamics of the index process, which tracks the flow of information between the reference and target. In the case of reversible PT, the index process retains no memory of the direction of travel, forcing the reference and target chain to communicate through a random walk. In contrast, for non-reversible PT, the index process maintains momentum in the direction of travel unless a rejection occurs, or a boundary is reached. We characterize communication efficiency through the round trip rate, measuring the number of round trips per unit time.

Our first contribution in Chapter 3 was to show that the index process is Markovian under the simplifying ELE assumption. We argue the round trip rate is inversely proportional to the expected round trip time for the index process, which is analytically tractable to compute (Theorem 1). It follows that non-reversible PT non-asymptotically dominates reversible PT in terms of round trip rates for any choice of N and annealing schedule $\mathcal{B}_N$ (Corollary 2).

Our first contribution in Chapter 3 was to show that the index process is Markovian under the simplifying ELE assumption. We argued that the round trip rate is inversely proportional to the expected round trip time for the index process, which is analytically tractable to compute (Theorem 1). It followed from Corollary 2 that non-reversible PT non-asymptotically dominates reversible PT in terms of round trip rates for any number of parallel chains, N, and the annealing schedule, $\mathcal{B}_N$.

We then studied PT in a novel asymptotic regime where we let N go to infinity through the

rejection rate. We defined the local communication barrier, $\lambda(\beta)$, as the instantaneous rejection rate, and the global communication barrier, $\Lambda = \int_0^1 \lambda(\beta)\mathrm{d}\beta$. The local communication barrier measures how rapidly the path of distributions π_β changes for small perturbations in β. We showed in Theorem 4 and Corollary 5 that the rejection rate between nearby chains approximates the integral of the local communication barrier, the sum of the rejection rate equals Λ with $O(N^{-2})$ error. The communication barrier is invariant to the schedule, and can be interpreted as a divergence between π_0 and π_1 measuring the difficulty of communicating along the path π_β.

In particular, we characterized the optimal round trip rate in terms of the global communication barrier Λ, which we used to identify an intrinsic limitation of both reversible and non-reversible PT (Section 3.2.4). Theorem 6 showed that as N increases to infinity, the round trip rate for reversible PT is asymptotically equivalent to $(2N + 2\Lambda)^{-1}$ in contrast to non-reversible PT, which increases with N but with diminishing returns to $(2 + 2\Lambda)^{-1}$. This demonstrates the optimal performance of PT decays with Λ, which accounts for the choice of a poor reference, dimension (Proposition 7), and multi-modality present in the target (Proposition 8).

We then combined the non-asymptotic and asymptotic analysis, to develop guidelines for tuning the schedule and number of chains for non-reversible PT to maximize the total number of round trips. We showed that the optimal annealing schedule satisfies a constant rejection rate across all chains. We also provided practical guidelines to optimally tune the schedule and efficiently allocate computation resources to maximize the performance of PT for general problems. This led to the development of Algorithm 5 which tunes PT, samples from the target, and computes the log-normalizing constants, given only the annealing path and computational budget as input.

Algorithm 5 also outputs the local and global communication barriers, which we used to develop diagnostic tools for practitioners to see the quality of their annealing path and schedule. By comparing the empirical round trip rate with $(2 + 2\Lambda)^{-1}$, we can also assess the efficiency of the implementation and the deviance from optimality. Before our methodology, there was no established criterion in the literature to measure the efficiency PT algorithms.

In Section 3.4 we applied Algorithm 5 to a diverse selection of 16 different models from statistics, physics, and biology to verify our theory and showcase the robustness of our methodology. In particular, our adaptive schedule tuning algorithm reliably converged for all models within ten adaptive rounds. We showed that the non-reversible PT is robust to violations in ELE, scales

to high-dimensional Bayesian inference problems with real data and multimodal posteriors. Our methodology outperformed state-of-the-art PT methods in terms of both round trips and ESS/second by a factor of 10-100x in all of our experiments.

A consequence of the analysis in Chapter 3 is that the optimal round trip rate that can be achieved is $(2 + 2\Lambda)^{-1}$. When Λ is large, it could mean a meagre round trip rate, even when tuned optimally with a large number of chains. One common situation this can occur is when π_0 and π_1 are nearly mutually singular, e.g. in the Bayesian setting where π_0 is a misspecified prior. The only way to improve the performance of PT is by decreasing Λ, which is a function of the annealing path. Traditionally in the PT literature, we presume the use of the "linear" annealing path $\pi_\beta \propto \pi_0^{1-\beta}\pi_1^{\beta}$ which linearly interpolates between π_0 and π_1 in log-scale. We abandoned this convention and developed a theory of PT for general, non-linear paths.

By generalizing the ELE assumptions from Chapter 3 to general paths, we showed that the non-asymptotic analysis of non-reversible PT extends to general non-linear annealing paths. We then expanded the definition of local and global communication barriers to non-linear paths and showed under mild regularity assumptions on the path; the asymptotic analysis developed in Chapter 3 applies to non-linear paths (Theorem 10). Given a family of annealing paths, we modified the tuning phase of Algorithm 5 to tune both the path as well as the schedule (Algorithm 6).

In Section 4.5 we analysed the geometry of annealing paths that take on values in a parametric family of distributions $\mathcal{M}$. We showed in Proposition 12 that the rejection rate induces a natural geometry on $\mathcal{M}$, where the local and global communication barrier can be interpreted as the speed the length of the annealing path, respectively. Hence, optimal tuning of PT algorithms is equivalent to finding a constant speed, length minimizing paths in $\mathcal{M}$, also known as geodesics. We showed that the geodesics with respect to geometry induced by PT are well-approximated by the geodesics generated by the Fisher information metric.

In Section 4.5.9, we demonstrated the potential of this framework when $\mathcal{M}$ is a location-scale family modeling the motivating example where the reference and target are nearly mutually singular. Here we showed $\mathcal{M}$ is a hyperbolic space with analytically tractable geodesics. We showed that the round trip rate induced by the geodesic path is exponentially larger compared to the naive translation path (Propostion 14).

In Section 4.6, we used the given reference and target to construct a 2-parameter exponential

annealing family $\mathcal{M}$ motivated by the characteristics of geodesics in local-scale families (Proposition 15). We also proposed a corresponding annealing path family using linear splines that can flexibly approximate any annealing path in $\mathcal{M}$. We showed that optimzing over this family is equivalent to find spline approximation to the geodesics in $\mathcal{M}$. We validated our theory using the Gaussian model and showed empirically we can surpass the theoretically optimal round trip rate for PT without path tuning.

In Chapter 5, we explained how such an innocuous algorithmic difference between reversible and non-reversible PT can lead to such a dramatic difference in performance. Given a regular annealing path and schedule generator, we computed the scaling limit of the index process for both reversible and non-reversible PT. In particular, we formally proved in Theorem 16(a) that by scaling the index process by N and speeding time by a factor or N^2, the index process for reversible PT converges to a Brownian motion on $[0, 1]$ with reflective boundary conditions independent of the annealing path and schedule (Theorem 16(a)). This diffusive scaling was presumed in the literature but without proof.

In contrast, we showed in Theorem 16(b) that for non-reversible PT by scaling the index process by a factor of N and speeding up time by N, the scaled index process weakly converges to a PDMP rather than a diffusion as previously presumed in the literature. The limit process is a scaled persistent random walk on $[0, 1]$ with reflective boundary conditions–the limit depends on the path and annealing schedule through the communication barrier and schedule generator. The limit process travels in a straight line and reverses direction according to the local communication barrier, and the global communication barrier is the total rate of reflection across $[0, 1]$. The inhomogeneity of the reflection rate is accounted for by the deviation from the optimal schedule.

6.2 Impact of work

MCMC methods were developed in conjunction with the first computers ever built (Robert and Casella, 2011), and their efficacy and popularity coincide with advancements in computational hardware. As parallel computing resources become cheaper, the demand for algorithms capable of efficiently utilizing them grows too. Our non-reversible PT framework was designed to be compatible with both existing and future MCMC methods and computing architectures. Its performance is

competitive with state-of-the-art sampling methods and will only improve with advancements in algorithmic and hardware developments.

Since our methodology does not make any structural assumptions on the state-space or target, it is particularly appealing to developers of probabilistic programming languages. Probabilistic programming is a programming paradigm where practitioners specify models in a high-level formalism, while the details of posterior inference are automated (van de Meent et al., 2018). Successful PPL implementations make state-of-the-art Monte Carlo methods accessible to general practitioners and dramatically increase their potential impact. Non-reversible PT can naturally be implemented in existing probabilistic programming languages that use MCMC and improve the quality of their inferences.

Our non-reversible PT framework has already been implemented as the default inference engine for Blang (Bouchard-Côté et al., 2021) for Bayesian inference on combinatorial spaces, and THEMIS (Broderick et al., 2020; Tiede, 2021) developed by physicists for Bayesian parameter estimation in astrophysics. Since our publication of Syed et al. (2021a), both Turing.jl (Ge et al., 2018), and TensorFlow Probability (Abadi et al., 2015) have updated their implementation of PT to use non-reversible communication and plan to implement our tuning guidelines in the near future.

Other research groups have also used our non-reversible PT framework and achieved high-performance multi-modal posterior exploration in other scientific disciplines. Most notably, the Event Horizon Telescope collaboration used THEMIS to reproduce the first image of Sagittarius A*, the supermassive black hole at the center of the galaxy M87 (Akiyama et al., 2019) (Figure 6.1 (Left)). Algorithm 5 from Chapter 3 was benchmarked against Vousden et al. (2016), a popular algorithm for tuning PT in the physics literature, with various local exploration kernels in Chapter 4 of Tiede (2021). They found samplers using Algorithm 5 converged faster compared with Vousden et al. (2016). More importantly, Algorithm 5 was the only one that reliably discovered all three dominant modes. In particular, Algorithm 5 in combination with NUTS (Hoffman et al., 2014) converged to the correct posterior 98% faster compared to THEMIS without non-reversible PT (Tiede, 2021, Chapter 4.6). The significant gains in performance and reliability THEMIS achieved from our non-reversible PT methodology were instrumental in modeling interstellar scattering in measurements (Issaoun et al., 2021). This research discovered magnetic polarization in the photograph (Akiyama et al., 2021) represented by the ripples in Figure 6.1 (Right).

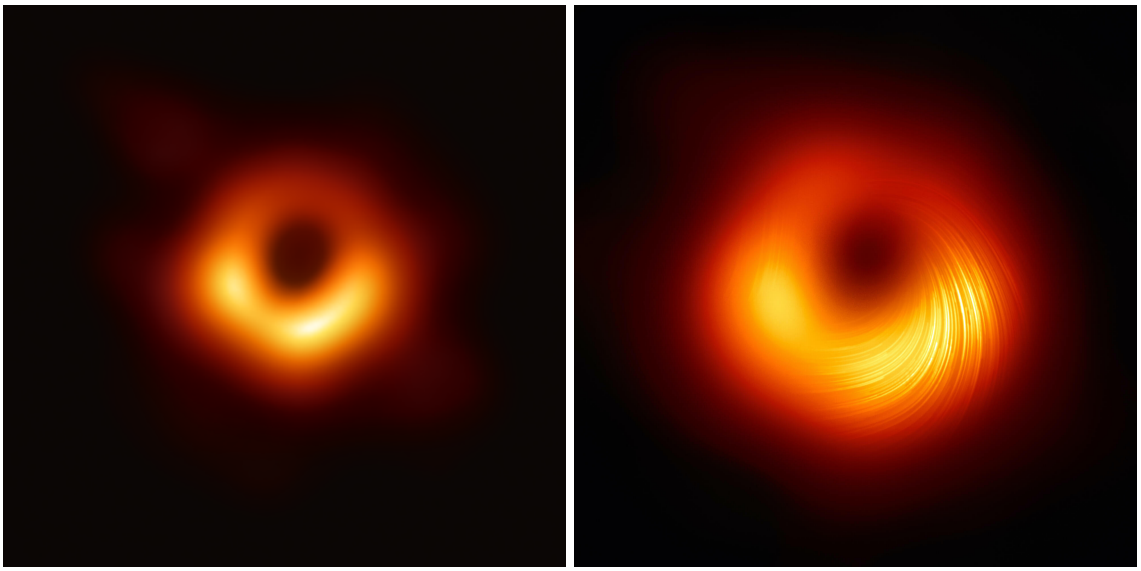

Figure 6.1: The first photograph of Sagittarius A*, the supermassive black hole at the center of galaxy M87 (left) and its update with Magnetic polarization (right) Source: `eventhorizontelescope.org`.

Our work was also used recently by Google research in conjunction with Hamiltonian Monte Carlo for solving Bayesian inference problems with poor conditioning and multi-modal posteriors (Langmore et al., 2021). Our non-reversible PT methodology was used by Dorri et al. (2020) to perform Bayesian phylogenetic tree reconstruction using low-depth genome-wide single-cell data. This was used to model time-series of single-cell cancer genomes (Salehi et al., 2021).

6.3 Future research directions

6.3.1 Weakening ELE assumption

The critical tool for our theoretical analysis was the ELE assumption which assumes the independence of the log-likelihood at each scan. We used ELE to study the index process for each machine as a Markov process and to make our theoretical analysis of the round trip rate tractable. The ELE is motivated by target distributions where multimodality arises from label switching and is satisfied if the local exploration kernel can draw independent samples within each mode. However, as discussed in Section 3.1.2, we do not expect the ELE to hold in practice. This implies the potential to develop a theory of non-reversible PT without ELE. We hypothesize we can retain theoretical guarantees if we weaken the ELE to assume the local exploration kernel is geometrically or uniformly ergodic on

each mode.

Weakening the ELE assumption would make the analysis more delicate since the index process would no longer be Markovian. However, it would open the possibility to incorporate information about the exploration kernel into designing better PT algorithms. It is worth exploring how the theory and tuning guidelines change when simultaneously tuning both PT and the exploration kernel. In particular, it would be of practical importance to know how PT interacts with popular local exploration kernels as Hamiltonian Monte Carlo (Langmore et al., 2021; Broderick et al., 2020; Tiede, 2021).

6.3.2 Mixing properties of round trips

The goal of PT is to improve mixing times and convergence rates of MCMC algorithms; however, our analysis optimizes the round trip rate rather than the spectral gap or ESS. The round trip rate measured the performance of the communication move in PT algorithms independent of the problem-specific local exploration kernel. This coarse grain approach to the PT analysis prevents us from making any rigorous claims about mixing in a particular instance of PT where the local exploration kernel does not satisfy the ELE condition.

Our empirical analysis shows that optimizing round trips improves mixing between modes and effective sample size, with gains in performance as the quality of the local exploration improves as seen in Figures 2.4 and 3.9. In particular, Figure 2.4 is compelling evidence that ESS is correlated with the total number of round trips and is worth further exploration. Understanding the mixing and rejuvenation structure of PT and determining how it relates to the round trip rates would be valuable for the theoretical understanding of PT algorithms and could give new insights on how to tune the local exploration kernel compatible with PT.

6.3.3 PT with variational reference

The motivation for Chapter 4 was to improve the performance of PT by reducing the communication barrier between π_0 and π_1. We did this by improving the path between them, to increase the overlap between successive distributions. Alternatively, we could have also chosen a better reference distribution to be "closer" to the target in terms of communication barrier.

Suppose $\mathcal{M}_0 = \{q_\theta : \theta \in \Theta\}$ is a flexible, easy to sample family of reference distributions over $\mathcal{X}$

that we can evaluate up to a constant. For example, $\mathcal{M}_0$ can be a family of Gaussian mixtures, mean-field approximations of π_1, or a family of distributions parametrized by a neural net. Using $\mathcal{M}_0$, we can construct an annealing path family $\mathcal{A} = \{\pi^\theta : \theta \in \Theta\}$, where $\pi_\beta^\theta \propto q_\theta^{1-\beta}\pi_1^\beta$ is the linear path between q_θ and π_1. We can then define the loss function, $\mathcal{L} : \mathcal{A} \to \mathbb{R}_+$, where $\mathcal{L}(\pi^\theta) = \Lambda(\pi^\theta)$ is the global communication barrier. We can improve the performance of PT by picking the path π minimizing $\mathcal{L}$. We can interpret the optimal reference π_0 as the projection of the target π_1 onto $\mathcal{M}_0$. This construction also naturally extends to path families with linear splines using the construction from Section 4.6.

If we use π_0 for $\pi \in \arg\min \mathcal{L}$ in place of the target, this becomes equivalent to doing variational inference (Blei et al., 2017) with the communication barrier as the objective rather than the Kullback–Leibler divergence. Variational inference using different divergences is an active area of research (Li and Turner, 2016; Regli and Silva, 2018; Masrani et al., 2019; Wan et al., 2020). We can use existing knowledge amassed by the variational inference community to improve the round trip rate for PT.

6.3.4 Geometric structure of annealing

We developed a geometric theory of PT induced by the rejection rate, where the local and global communication barriers correspond to the speed and length respectively. This allows us to reinterpret the tuning of PT algorithms as computing a geodesic. In particular, it would be valuable to characterize the geodesics and how curvature relates to designing better annealing paths. It is worth understanding the geometry of the exponential annealing family from Section 4.6.1 and its connection to information geometry more generally. We already saw from Proposition 7 in the high dimension scaling limit, the Finsler structure for PT coincides with the Fisher information metric. This connection to information geometry opens up the possibility to study PT using more sophisticated tools from differential geometry. Conversely, PT can be used as a tool for the field of applied information geometry (Amari, 2016) since Algorithm 6 approximates the geodesics for any specified annealing family $\mathcal{M}$ viewed as an information manifold. We could see applications of Algorithm 6 potentially outside of the MCMC context.

6.3.5 Beyond PT

This thesis showed that non-reversible PT is competitive with state-of-the-art Monte Carlo methods. Still, it remains an open problem to determine when practitioners should use non-reversible PT versus another class of successful methods such as SMC (Del Moral et al., 2006). The geometric framework developed in Chapter 4 extends naturally beyond PT to any annealing-based algorithm where the efficiency is measured using a regular divergence. Examples include AIS (Neal, 2001), SMC, and optimal transport (Villani, 2009), where the Finsler structure is induced by the KL divergence (Grosse et al., 2013), χ^2 divergence (Agapiou et al., 2017; Chatterjee and Diaconis, 2018), and the squared Wasserstein distance (Lott and Villani, 2009; Figalli and Villani, 2011) respectively.

Our geometric framework developed in Chapter 4 unifies these seemingly disparate methodologies, each constituting a rich literature. There is potential to use the tools and insights from PT to study these other algorithms. In particular, we can apply our theory and guidelines developed for PT to create a new approach to tuning for other methods. Conversely, the tools from these areas can potentially lead to novel advancements in PT both theoretically and algorithmically.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Akiyama, K., Alberdi, A., Alef, W., Asada, K., Azulay, R., Baczko, A.-K., Ball, D., Baloković, M., Barrett, J., Bintley, D., et al. (2019). First M87 event horizon telescope results. IV. Imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4.

Akiyama, K., Algaba, J. C., Alberdi, A., Alef, W., Anantua, R., Asada, K., Azulay, R., Baczko, A.-K., Ball, D., Baloković, M., et al. (2021). First M87 event horizon telescope results. VII. Polarization of the ring. *The Astrophysical Journal Letters*, 910(1):L12.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505.

Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568.

Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J., and Theis, F. J. (2017). Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Systems Biology*, 11(1):63.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557.

Baxter, R. J. (2007). *Exactly Solved Models in Statistical Mechanics*. Dover books on Physics. Dover Publications.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534.

Bierkens, J., Bouchard-Côté, A., Doucet, A., Duncan, A. B., Fearnhead, P., Lienart, T., Roberts, G. O., and Vollmer, S. J. (2018). Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statistics & Probability Letters*, 136:148–154.

Bierkens, J. and Roberts, G. O. (2017). A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model. *The Annals of Applied Probability*, 27(2):846–882.

Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons.

Bishop, R. L. (2013). Riemannian geometry. *arXiv preprint arXiv:1303.5390*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Böttcher, B., Schilling, R., and Wang, J. (2013). Lévy matters. III, volume 2099 of Lecture Notes in Mathematics.

Bouchard-Côté, A., Chern, K., Cubranic, D., Hosseini, S., Hume, J., Lepur, M., Ouyang, Z., and Sgarbi, G. (2021). Blang: Probabilitistic programming for combinatorial spaces. *Journal of Statistical Software*, (Accepted).

Brekelmans, R., Masrani, V., Bui, T., Wood, F., Galstyan, A., Steeg, G. V., and Nielsen, F. (2020). Annealed importance sampling with q-paths. *arXiv preprint arXiv:2012.07823*.

Bridson, M. R. and Haefliger, A. (2013). *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media.

Brockwell, A. E. (2006). Parallel Markov chain Monte Carlo simulation by pre-fetching. *Journal of Computational and Graphical Statistics*, 15(1):246–261.

Broderick, A. E., Gold, R., Karami, M., Preciado-López, J. A., Tiede, P., Pu, H.-Y., Akiyama, K., Alberdi, A., Alef, W., Asada, K., et al. (2020). THEMIS: A parameter estimation framework for the event horizon telescope. *The Astrophysical Journal*, 897(2):139.

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413.

Chandra, R., Jain, K., Deo, R. V., and Cripps, S. (2019). Langevin-gradient parallel tempering for Bayesian neural learning. *Neurocomputing*, 359:315–326.

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135.

Chen, F., Lovász, L., and Pak, I. (1999). Lifting Markov chains to speed up mixing. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 275–281. ACM.

Chen, J., Lu, D., Xiu, Z., Bai, K., Carin, L., and Tao, C. (2021). Variational inference with Holder bounds. *arXiv preprint arXiv:2111.02947*.

Cho, K., Raiko, T., and Ilin, A. (2010). Parallel tempering is efficient for learning restricted Boltzmann machines. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE.

Costa, S. I., Santos, S. A., and Strapasson, J. E. (2015). Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318.

Dabak, A. G. and Johnson, D. H. (2002). Relations between Kullback-Leibler distance and Fisher information.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, 84(408):945–957.

Davidson-Pilon, C. (2015). *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley Professional, New York, 1st edition.

Davis, M. H. (1993). *Markov Models & Optimization*. Chapman and Hall.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(3):411–436.

Deligiannidis, G., Paulin, D., Bouchard-Côté, A., and Doucet, A. (2018). Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *arXiv preprint arXiv:1808.04299*.

Desjardins, G., Luo, H., Courville, A., and Bengio, Y. (2014). Deep tempering. *arXiv preprint arXiv:1410.0123*.

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

Diaz, A., Argüelles, C., Collin, G., Conrad, J., and Shaevitz, M. (2020). Where are we with light sterile neutrinos? *Physics Reports*, 884:1–59.

Dorri, F., Salehi, S., Chern, K., Funnell, T., Williams, M., Lai, D., Andronescu, M., Campbell, K. R., McPherson, A., Aparicio, S., Roth, A., Shah, S. P., and Bouchard-Côté, A. (2020). Efficient Bayesian inference of phylogenetic trees from large scale, low-depth genome-wide single-cell data. *bioR Xiv 2020.05.06.058180*.

Dupuis, P., Liu, Y., Plattner, N., and Doll, J. D. (2012). On the infinite swapping limit for parallel tempering. *SIAM Multiscale Modeling & Simulation*, 10(3):986–1022.

Ethier, S. N. and Kurtz, T. G. (2009). *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons.

Fang, Y., Feng, S., Tam, K.-M., Yun, Z., Moreno, J., Ramanujam, J., and Jarrell, M. (2014). Parallel tempering simulation of the three-dimensional Edwards-Anderson model with compact asynchronous multispin coding on GPU. *Computer Physics Communications*, 185(10):2467–2478.

Figalli, A. and Villani, C. (2011). Optimal transport and curvature. In *Nonlinear PDE's and Applications*, pages 171–217. Springer.

Friesen, J. (2015). *Java Threads and the Concurrency Utilities*. Apress, Berkeley, CA, USA, 1st edition.

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

Ge, H., Xu, K., and Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Interface Proceedings*.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. (2013). Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*.

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

Hayasaka, K., Gojobori, T., and Horai, S. (1988). Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, 5(6):626–644. Publisher: Oxford Academic.

Hind, P. (2019). RMS Titanic passenger dataset. data retrieved from `https://tinyurl.com/y55c8kc7`.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Hsieh, Y.-D., Kao, Y.-J., and Sandvik, A. W. (2013). Finite-size scaling method for the Berezinskii-Kosterlitz-Thouless transition. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09001.

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

Hukushima, K. and Sakai, Y. (2013). An irreversible Markov-chain Monte Carlo method with skew detailed balance conditions. 473(1):012012.

Issaoun, S., Johnson, M. D., Blackburn, L., Broderick, A., Tiede, P., Wielgus, M., Doeleman, S. S., Falcke, H., Akiyama, K., Bower, G. C., Brinkerink, C. D., Chael, A., Cho, I., Gómez, J. L., Hernández-Gómez, A., Hughes, D., Kino, M., Krichbaum, T. P., Liuzzo, E., Loinard, L., Markoff, S., Marrone, D. P., Mizuno, Y., Moran, J. M., Pidopryhora, Y., Ros, E., Rygl, K., Shen, Z.-Q., and Wagner, J. (2021). Persistent non-Gaussian structure in the image of Sagittarius A* at 86 GHz. *The Astrophysical Journal*, 915(2):99.

Jacka, S. and Hernández-Hernández, M. E. (2019). Minimising the expected commute time. *Stochastic Processes and their Applications*.

Jacob, P., Robert, C. P., and Smith, M. H. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics*, 20(3):616–635.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer, 2nd edition.

Kamberaj, H. (2020). *Molecular Dynamics Simulations in Statistical Physics: Theory and Applications*. Springer.

Katzgraber, H. G., Trebst, S., Huse, D. A., and Troyer, M. (2006). Feedback-optimized parallel tempering Monte Carlo. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(03):P03018.

Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *The Journal of chemical physics*, 3(5):300–313.

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

Kone, A. and Kofke, D. A. (2005). Selection of temperature intervals for parallel-tempering simulations. *The Journal of Chemical Physics*, 122(20):206101.

Łacki, M. K. and Miasojedow, B. (2016). State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Statistics and Computing*, 26(5):951–964.

Langmore, I., Dikovsky, M., Geraedts, S., Norgaard, P., and von Behren, R. (2021). Hamiltonian Monte Carlo in inverse problems; ill-conditioning and multi-modality. *arXiv preprint arXiv:2103.07515*.

Lee, A., Yau, C., Giles, M. B., Doucet, A., and Holmes, C. C. (2010). On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of computational and graphical statistics*, 19(4):769–789.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Leiserson, C. E., Schardl, T. B., and Sukha, J. (2012). Deterministic parallel random-number generation for dynamic-multithreading platforms. *MIT web domain*.

Lelièvre, T., Stoltz, G., and Rousset, M. (2010). *Free Energy Computations: A Mathematical Perspective*. World Scientific.

Leonhardt, C., Schwake, G., Stögbauer, T. R., Rappl, S., Kuhr, J.-T., Ligon, T. S., and Rädler, J. O. (2014). Single-cell mRNA transfection studies: delivery, kinetics and statistics by numbers. *Nanomedicine: Nanotechnology, Biology, and Medicine*, 10(4):679–688.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. *arXiv preprint arXiv:1602.02311*.

Lingenheil, M., Denschlag, R., Mathias, G., and Tavan, P. (2009). Efficiency of exchange schemes in replica exchange. *Chemical Physics Letters*, 478(1-3):80–84.

Lott, J. and Villani, C. (2009). Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pages 903–991.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Masoliver, J., Porra, J. M., and Weiss, G. H. (1992). Solutions of the telegrapher's equation in the presence of traps. *Physical Review A*, 45(4):2222.

Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Steeg, G. V., and Wood, F. (2021). q-Paths: Generalizing the geometric annealing path using power means. *arXiv preprint arXiv:2107.00745*.

Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. *arXiv preprint arXiv:1907.00031*.

McDowell, J. (2019). Launch logs. data retrieved from `https://tinyurl.com/y5veq9yf`.

McGrayne, S. B. (2011). *The theory that would not die*. Yale University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, 44(247):335–341.

Miasojedow, B., Moulines, E., and Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664.

Mingas, G. and Bouganis, C.-S. (2012). Parallel tempering MCMC acceleration using reconfigurable hardware. In Choy, O. C. S., Cheung, R. C. C., Athanas, P., and Sano, K., editors, *Reconfigurable Computing: Architectures, Tools and Applications*, Lecture Notes in Computer Science, pages 227–238. Springer Berlin Heidelberg.

Müller, N. F. and Bouckaert, R. R. (2020). Adaptive parallel tempering for BEAST 2. *bioRxiv*.

Nadler, W. and Hansmann, U. H. E. (2007). Dynamics and optimal number of replicas in parallel tempering simulations. *Physical Review E*, 76(6):065701.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Neklyudov, K., Egorov, E., Shvechikov, P., and Vetrov, D. (2018). Metropolis-hastings view on variational inference and adversarial training. *arXiv preprint arXiv:1810.07151*.

Nielsen, F. (2013). Cramér-Rao lower bound and information geometry. In *Connected at Infinity II*, pages 18–37. Springer.

Nielsen, F. (2020). An elementary introduction to information geometry. *Entropy*, 22(10):1100.

Nielsen, F. (2021). On information projections between multivariate elliptical and location-scale families. *arXiv preprint arXiv:2101.03839*.

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

Polyanskiy, Y. (2020). Information theoretic methods in statistics and computer science: f-divergences. `http://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf`.

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

Rathore, N., Chopra, M., and de Pablo, J. J. (2005). Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics*, 122(2):024111.

Regli, J.-B. and Silva, R. (2018). Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*.

Rischard, M., Jacob, P. E., and Pillai, N. (2018). Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *arXiv preprint arXiv:1810.01382*.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, pages 102–115.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367.

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

Sakai, Y. and Hukushima, K. (2016). Irreversible simulated tempering. *Journal of the Physical Society of Japan*, 85(10):104002.

Salehi, S., Kabeer, F., Ceglia, N., Andronescu, M., Williams, M. J., Campbell, K. R., Masud, T., Wang, B., Biele, J., Brimhall, J., Gee, D., Lee, H., Ting, J., Zhang, A. W., Tran, H., Flanagan, C. O., Dorri, F., Rusk, N., de Algara, T. R., Lee, S. R., Cheng, B. Y. C., Eirew, P., Kono, T., Pham, J., Grewal, D., Lai, D., Moore, R., Mungall, A. J., Marra, M. A., Consortium, I., McPherson, A., Bouchard-Côté, A., Aparicio, S., and Shah, S. P. (2021). Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature*, 595:585–590.

Saranen, J. and Seikkala, S. (1988). Solution of a nonlinear two-point boundary value problem with Neumann-type boundary data. *Journal of Mathematical Analysis and Applications*, 135(2):691–701.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Shen, Z. (2006). Riemann-Finsler geometry with applications to information geometry. *Chinese Annals of Mathematics, Series B*, 27(1):73–94.

Steele, G. and Lea, D. (2013). Splittable random application programming interface. `https://docs.oracle.com/javase/8/docs/api/java/util/SplittableRandom.html`. [Online; accessed 6-May-2019].

Swendsen, R. H. and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2021a). Non-reversible parallel tempering: A scalable highly parallel MCMC scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Syed, S., Romaniello, V., Campbell, T., and Bouchard-Cote, A. (2021b). Parallel tempering on optimized paths. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10033–10042. PMLR.

Tamássy, L. (2005). Finsler spaces corresponding to distance spaces. In *Proc. of the Conf., Contemporary Geometry and Related Topics, Belgrade, Serbia and Montenegro, June*, pages 485–495. Citeseer.

Tataru, P., Simonsen, M., Bataillon, T., and Hobolth, A. (2017). Statistical inference in the Wright-Fisher model using allele frequency data. *Systematic Biology*, 66(1):e30–e46.

Tawn, N. G., Roberts, G. O., and Rosenthal, J. S. (2020). Weight-preserving simulated tempering. *Statistics and Computing*, 30(1):27–41.

Tiede, P. (2021). *The Nature and Impact of Active Galactic Nuclei*. PhD dissertation, University of Waterloo.

Turitsyn, K. S., Chertkov, M., and Vucelja, M. (2011). Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414.

Van de Meent, J.-W., Paige, B., Yang, H., and Wood, F. (2018). An introduction to probabilistic programming. *arXiv preprint arXiv:1809.10756*.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vousden, W., Farr, W. M., and Mandel, I. (2016). Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Monthly Notices of the Royal Astronomical Society*, 455(2):1919–1937.

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

Wan, N., Li, D., and Hovakimyan, N. (2020). f-Divergence variational inference. *Advances in Neural Information Processing Systems*, 33.

Wang, X., Guo, F., Heller, K. A., and Dunson, D. B. (2015). Parallelizing MCMC with random partition trees. *arXiv preprint arXiv:1506.03164*.

Whitfield, T., Bu, L., and Straub, J. (2002). Generalized parallel sampling. *Physica A: Statistical Mechanics and its Applications*, 305(1-2):157–171.

Woodard, D. B., Schmidler, S. C., Huber, M., et al. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640.

Wu, C. and Robert, C. P. (2017). Average of recentered parallel MCMC for big data. *arXiv preprint arXiv:1706.04780*.

Wu, F. (2017). Irreversible Parallel Tempering and an Application to a Bayesian Nonparametric Latent Feature Model. Master's thesis, Oxford University.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic biology*, 60(2):150–160.

Zhou, Y., Johansen, A. M., and Aston, J. A. (2016). Toward automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726.

Zhu, P., Bouchard-Côté, A., and Campbell, T. (2020). Slice sampling for general completely random measures. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 36.

Zimmermann, H., Wu, H., Esmaeili, B., and van de Meent, J.-W. (2021). Nested variational inference. *arXiv preprint arXiv:2106.11302*.

Appendix A

Technical Proofs

A.1 Chapter 3

A.1.1 Theorem 1

Proof of Theorem 1. To simplify notation for the rest of the proof, let $T_\uparrow$ and $T_\downarrow$ be the hitting times to the target and reference defined by,

$$T_\uparrow = \min\{t : (I_t, \epsilon_t) = (N, 1)\}, \quad T_\downarrow = \min\{t > T_\uparrow : (I_t, \epsilon_t) = (0, -1)\}.$$

We will also denote

$$s_n = s^{(n-1,n)}, \qquad r_n = 1 - s^{(n-1,n)}.$$

Expected round trip times for SEO

If we define $a_\bullet^n = \mathbb{E}_{\mathrm{SEO}}(T_\bullet | I_0 = n)$ for $n = 0, \ldots, N$ and $\bullet \in \{\uparrow, \downarrow\}$, then we have

$$\mathbb{E}_{\mathrm{SEO}}(T) = a_\uparrow^0 + a_\downarrow^N. \tag{A.1}$$

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

$$a_\bullet^n = \frac{1}{2} s_{n+1}(a_\bullet^{n+1} + 1) + \frac{1}{2} s_n (a_\bullet^{n-1} + 1) + \frac{1}{2}(r_{n+1} + r_n)(a_\bullet^n + 1). \tag{A.2}$$

For $n = 1, \ldots, N$, we substitute in $b^n_\bullet = a^n_\bullet - a^{n-1}_\bullet$ into (A.2). After simplification, $b^n_\bullet$ satisfies the following recursive relation

$$-2 = s_{n+1} b^{n+1}_\bullet - s_n b^n_\bullet. \tag{A.3}$$

The solutions to (A.3) are

$$s_n b^n_\bullet = s_1 b^1_\bullet - 2(n-1), \tag{A.4}$$

or equivalently

$$s_n b^n_\bullet = s_N b^N_\bullet + 2(N-n). \tag{A.5}$$

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

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

$$a^0_\uparrow = \frac{1}{2} s_1 (a^1_\uparrow + 1) + \left(1 - \frac{1}{2} s_1\right)(a^0_\uparrow + 1),$$

or equivalently

$$s_1 b^1_\uparrow = -2.$$

Substituting this into (A.4) implies $s_n b^n_\uparrow = -2n$. By summing $b^n_\uparrow = a^n_\uparrow - a^{n-1}_\uparrow$ from $n = 1, \ldots, N$ and, noting $a^N_\uparrow = 0$, we get

$$a^0_\uparrow = \sum_{n=1}^{N} \frac{2n}{s_n}. \tag{A.6}$$

Computation of $a_\downarrow^N$. Similarly to determine $a_\downarrow^N$ we note that 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

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

or equivalently

$$s_N b_\downarrow^N = 2.$$

Substituting this into (A.5) implies $s_n b_\downarrow^n = 2 + 2(N-n)$. By summing $b_\downarrow^n = a_\downarrow^n - a_\downarrow^{n-1}$ from $n = 1, \ldots, N$ and, noting $a_\downarrow^0 = 0$, we get

$$a_\downarrow^N = \sum_{n=1}^{N} \frac{2(N-n)+2}{s_n}. \tag{A.7}$$

Substituting in (A.6) and (A.7) into (A.1), it follows that

$$\begin{aligned}
\mathbb{E}_{\mathrm{SEO}}(T) &= \sum_{n=1}^{N} \frac{2n}{s_n} + \sum_{n=1}^{N} \frac{2(N-n)+2}{s_n} \\
&= 2(N+1)\sum_{n=1}^{N} \frac{1}{s_n} \\
&= 2N(N+1) + 2(N+1)\sum_{n=1}^{N} \frac{r_n}{s_n}.
\end{aligned}$$

Expected round trip times for DEO

If we define $a_\bullet^{n,\epsilon} = \mathbb{E}_{\mathrm{DEO}}(T_\bullet | I_0 = n, \epsilon_0 = \epsilon)$ for $n = 0, \ldots, N$, $\epsilon \in \{+,-\}$ and $\bullet \in \{\uparrow, \downarrow\}$, then we have

$$\mathbb{E}_{\mathrm{DEO}}(T) = a_\uparrow^{0,-} + a_\downarrow^{N,+}. \tag{A.8}$$

Note that for $n = 1, \ldots, N-1$ $a_\bullet^{n,\epsilon}$ satisfies the recursion relations

$$a_\bullet^{n,+} = s_{n+1}(a_\bullet^{n+1,+} + 1) + r_{n+1}(a_\bullet^{n,-} + 1), \tag{A.9}$$

$$a_\bullet^{n,-} = s_n(a_\bullet^{n-1,-} + 1) + r_n(a_\bullet^{n,+} + 1). \tag{A.10}$$

If we substitute $c_\bullet^n = a_\bullet^{n,+} + a_\bullet^{n-1,-}$, and $d_\bullet^n = a_\bullet^{n,+} - a_\bullet^{n-1,-}$ into (A.9) and (A.10) and simplify, we obtain

$$a_\bullet^{n+1,+} - a_\bullet^{n,+} = r_{n+1}d_\bullet^{n+1} - 1, \tag{A.11}$$

$$a_\bullet^{n,-} - a_\bullet^{n-1,-} = r_n d_\bullet^n + 1. \tag{A.12}$$

By subtracting and adding (A.11) and (A.12), we obtain a joint recursion relation for $c_\bullet^n$ and $d_\bullet^n$ of the form

$$c_\bullet^{n+1} - c_\bullet^n = r_{n+1}d_\bullet^{n+1} + r_n d_\bullet^n, \tag{A.13}$$

$$d_\bullet^{n+1} - d_\bullet^n = r_{n+1}d_\bullet^{n+1} + r_n d_\bullet^n - 2. \tag{A.14}$$

Note that (A.14) can be rewritten as

$$s_{n+1}d_\bullet^{n+1} - s_n d_\bullet^n = -2. \tag{A.15}$$

Once one has expressions for $c_\bullet^n$ and $d_\bullet^n$, then we can recover $a_\bullet^{n,\epsilon}$ by using

$$a_\bullet^{n,+} = \frac{c_\bullet^n + d_\bullet^n}{2},$$
$$a_\bullet^{n-1,-} = \frac{c_\bullet^n - d_\bullet^n}{2}.$$

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

Computation of $a_\uparrow^{0,-}$. Note that $a_\uparrow^{0,-} = a_\uparrow^{0,+} + 1$. We can substitute this into (A.11) to get $s_1 d_\uparrow^1 = -2$, which combined with (A.15) implies

$$s_n d_\uparrow^n = -2n. \tag{A.16}$$

Since $a_\uparrow^{N,+} = 0$ we have $c_\uparrow^N = -d_\uparrow^N$, so by summing (A.13) we get

$$\begin{aligned} 2a_\uparrow^{0,-} &= c_\uparrow^1 - d_\uparrow^1 \\ &= c_\uparrow^N - d_\uparrow^1 - \sum_{n=1}^{N-1} (c_\uparrow^{n+1} - c_\uparrow^n) \\ &= -d_\uparrow^N - d_\uparrow^1 - \sum_{n=1}^{N-1} (r_{n+1} d_\uparrow^{n+1} + r_n d_\uparrow^n) \\ &= -s_N d_\uparrow^N - s_1 d_\uparrow^1 - 2\sum_{n=1}^{N} r_n d_\uparrow^n. \end{aligned} \tag{A.17}$$

After substituting (A.16) into (A.17), we obtain

$$a_\uparrow^{0,-} = N + 1 + \sum_{n=1}^{N} \frac{2nr_n}{s_n}. \tag{A.18}$$

Computation of $a_\downarrow^{N,+}$. Note that $a_\downarrow^{N,+} = a_\uparrow^{N,-} + 1$. We can substitute this expression into (A.12) to get $s_N d_\downarrow^N = 2$, which combined with (A.15) implies

$$s_n d_\downarrow^n = 2(N - n + 1). \tag{A.19}$$

Since $a_\downarrow^{0,-} = 0$ we have $c_\downarrow^1 = d_\downarrow^1$, so by summing (A.13) we get

$$\begin{aligned} 2a_\downarrow^{N,+} &= c_\downarrow^N + d_\downarrow^N \\ &= c_\downarrow^1 + d_\downarrow^N + \sum_{n=1}^{N-1} (c_\downarrow^{n+1} - c_\downarrow^n) \\ &= d_\downarrow^1 + d_\downarrow^N + \sum_{n=1}^{N-1} (r_{n+1} d_\downarrow^{n+1} + r_i d_\downarrow^n) \\ &= s_1 d_\downarrow^1 + s_N d_\downarrow^N + 2\sum_{n=1}^{N} r_n d_\downarrow^n. \end{aligned} \tag{A.20}$$

After substituting in (A.19) into (A.20), we obtain

$$a_{\downarrow}^{N,+} = N + 1 + \sum_{n=1}^{N} \frac{2(N-n+1)r_n}{s_n}. \tag{A.21}$$

Finally, by substituting (A.18) and (A.21) into (A.8), it follows that

$$\mathbb{E}_{\mathrm{DEO}}[T] = 2(N+1) + 2(N+1)\sum_{n=1}^{N} \frac{r_n}{s_n}.$$

□

A.1.2 Theorem 3

Proof of Theorem 3. We first want to verify that $\mathbb{E}_{\bar{p}}[|\Delta W|^3] < \infty$. This follows from the integrability of ΔW^3 with respect to p, p' and the arithmetic and geometric means inequality

$$\bar{p}(x) \propto \sqrt{p(x)p'(x)} \leq \frac{1}{2}(p(x) + p'(x)).$$

Let us define $\lambda_k = 2^{-k}\mathbb{E}_{\bar{p}}[|\Delta W(X') - \Delta W(X)|^k]$ for $X, X' \sim \bar{p}$. By applying Taylor's remainder theorem to the numerator and denominator in Equation (3.12) to the third order we get

$$r(p, p') = 1 - \frac{1 - \lambda_1 + \lambda_2 + R'}{1 + \lambda_2 + R''} \tag{A.22}$$

where the remainders R', R'' satisfy

$$|R'|, |R''| \leq C'\mathbb{E}_{\bar{p}}[|\Delta W|^3]$$

for some constant $C' > 0$. Only the even terms remain in the denominator since $\mathbb{E}[(\Delta W(X') - \Delta W(X))^k] = 0$ for k odd by symmetry. By applying Taylor's remainder theorem again to (A.22) for the function $(1+x)^{-1}$ we get that the λ_2 terms cancel and we are left with

$$r(p, p') = \lambda_1 + R,$$

where the remainder R satisfies

$$|R| \leq C\mathbb{E}_{\bar{p}}[|\Delta W|^3]$$

for some finite constant C. □

A.1.3 Theorem 4

Lemma 21. *Let $f : \mathcal{X} \to \mathbb{R}$ be measurable. Then if $E_0[|f|], \mathbb{E}_1[|f|] < \infty$, then*

$$\|f\|_\pi = \sup_{\beta \in [0,1]} \mathbb{E}_\beta[|f|] < \infty. \tag{A.23}$$

Proof of Lemma 21. By the weighted AM-GM inequality for all $\beta \in [0, 1]$,

$$\pi_0(x)^{1-\beta}\pi_1(x)^\beta \leq (1-\beta)\pi_0(x) + \beta\pi_1(x).$$

By integrating over $\mathcal{X}$,

$$\begin{aligned}
\pi_\beta(|f|) &= \frac{1}{Z(\beta)} \int_{\mathcal{X}} |f(x)|\pi_0(x)^{1-\beta}\pi_1(x)^\beta \mathrm{d}x \\
&\leq (1-\beta)\frac{1}{Z(\beta)} \int_{\mathcal{X}} |f(x)|\pi_0(x)\mathrm{d}x + \beta\frac{1}{Z(\beta)} \int_{\mathcal{X}} |f(x)|\pi_1(x)\mathrm{d}x \\
&= (1-\beta)\frac{Z(0)}{Z(\beta)}\mathbb{E}_0[|f|] + \beta\frac{Z(1)}{Z(\beta)}\mathbb{E}_1[|f|] < \infty.
\end{aligned}$$

Since $Z(\beta)$ is continuous and positive on $[0, 1]$, (A.23) follows from the extreme value theorem. □

Lemma 22. *Suppose $\mathbb{E}_0[|V|^3], \mathbb{E}_0[|V|^3] < \infty$ then $\lambda \in C^2([0, 1])$ with derivatives satisfying,*

$$\left\| \frac{d^2\lambda}{d\beta^2} \right\|_\infty \leq C\|V^3\|_\pi, \tag{A.24}$$

where $\|V^3\|_\pi = \sup_\beta \mathbb{E}_\beta[|V|^3]$.

Proof. Suppose V^k is integrable with respect to π_0 and π_1, we want to show here that $\lambda : [0, 1] \to \mathbb{R}_+$

given by

$$\lambda(\beta) = \frac{1}{2}\int_{\mathcal{X}^2} |V(x) - V(y)|\pi_\beta(x)\pi_\beta(y)\mathrm{d}x\mathrm{d}y = \frac{g(\beta)}{2Z(\beta)^2}, \tag{A.25}$$

where $Z, g : [0,1] \to \mathbb{R}_+$ satisfy,

$$\begin{aligned} Z(\beta) &= \int_{\mathcal{X}} \tilde{\pi}_\beta(x)\mathrm{d}x, \\ g(\beta) &= \int_{\mathcal{X}^2} |V(x) - V(y)|\tilde{\pi}_\beta(x)\tilde{\pi}_\beta(y)\mathrm{d}x\mathrm{d}y, \end{aligned}$$

where $\tilde{\pi}_\beta(x) = \exp(W_\beta)$ is the un-normalized density of π_β. Since $Z(\beta) > 0$ on $[0,1]$, if we can show that $Z, g \in C^2([0,1])$ then it implies that $\lambda \in C^2([0,1])$.

Regularity of Z: Note that $\tilde{\pi}_\beta(x)$ satisfies

$$\frac{\partial^j \tilde{\pi}_\beta}{\partial \beta^j}(x) = V(x)^j \tilde{\pi}_\beta(x).$$

For all $j \leq 3$,

$$\sup_{\beta\in[0,1]} \left| \frac{\partial^j \tilde{\pi}_\beta}{\partial \beta^j}(x) \right| \leq |V(x)|^j \pi_0(x) + |V(x)|^j \pi_1(x). \tag{A.26}$$

The bound in (A.26) is uniform in β and is integrable. By Leibniz integration rule we have $Z \in C^3([0,1])$ with derivatives satisfying

$$\frac{d^j Z}{d\beta^j} = Z(\beta) \int_{\mathcal{X}} V(x)^j \pi_\beta(x)\mathrm{d}x.$$

Since $Z(\beta)$ is continuous on $[0,1]$, we have

$$\left\| \frac{d^j Z}{d\beta^j} \right\|_\infty \leq \|Z\|_\infty \|V^j\|_\pi. \tag{A.27}$$

where $\|V^j\|_\pi = \sup_\beta \mathbb{E}_\beta[|V|^j]$.

Regularity of g: Let $h(x,y,\beta) = |V(x) - V(y)|\tilde{\pi}_\beta(x)\tilde{\pi}_\beta(y)$. The partial derivatives satisfy,

$$\frac{\partial^j}{\partial\beta^j}h(x,y,\beta) = (-1)^j|V(x) - V(y)|(V(x) + V(y))^j\tilde{\pi}_\beta(x)\tilde{\pi}_\beta(y).$$

Similar to (a), we have for all $\beta \in [0,1]$, $j \leq 2$,

$$\begin{aligned}\sup_{\beta\in[0,1]}\left|\frac{\partial^j}{\partial\beta^j}h(x,y,\beta)\right| \leq\ & |V(x) - V(y)||V(x) + V(y)|^j\pi_0(x)\pi_0(y) \\ & + |V(x) - V(y)||V(x) + V(y)|^j\pi_1(x)\pi_1(y). \qquad \text{(A.28)}\end{aligned}$$

The left hand side of (A.28) dominates $\frac{\partial^j h}{\partial\beta^j}$ uniformly in β. It is integrable by Lemma 21 and using the fact that V^k is integrable with respect to π_0 and π. By the Leibniz integration rule,

$$\frac{d^j g}{d\beta^j} = Z(\beta)^2\int_{\mathcal{X}^2}(-1)^j|V(x) - V(y)|(V(x) + V(y))^j\pi_\beta(x)\pi_\beta(y)\mathrm{d}x\mathrm{d}y.$$

Again by the continuity of $Z(\beta)$ we have for $j \leq 2$.

$$\left\|\frac{d^j g}{d\beta^j}\right\|_\infty \leq \|Z\|_\infty^2\|V^j\|_\pi. \qquad \text{(A.29)}$$

Finally we get (A.24) by applying the quotient rule to (A.25) then using (A.27), (A.29). □

Proof of Theorem 4. We first note that the log-likelihood ratio between $\pi_\beta \propto \exp(W_\beta), \pi_{\beta'} \propto \exp(W_{\beta'})$ satisfies,

$$\Delta W = W_{\beta'} - W_\beta = (\beta' - \beta)V(x).$$

Applying Theorem 3 to π_β and $\pi_{\beta'}$ we get for some $C' > 0$

$$\left|r(\beta,\beta') - (\beta' - \beta)\lambda(\bar{\beta})\right| \leq C'\mathbb{E}_{\bar{\beta}}[|V|^3]|\beta' - \beta|^3, \qquad \text{(A.30)}$$

where $\bar{\beta} = (\beta + \beta')/2$ and $\lambda(\bar{\beta})$ is defined by (3.13). It follows from Lemma 21 that

$$\|V^3\|_\pi = \sup_{\beta\in[0,1]}\mathbb{E}_\beta[|V|^3] < \infty.$$

Notice that $(\beta' - \beta)\lambda(\bar{\beta})$ is the Riemann sum for $\int_\beta^{\beta'} \lambda(u)\mathrm{d}u$ with a single rectangle. By Lemma 22, we have λ is twice continuously differentiable and thus the standard midpoint rule error estimates yields

$$\left| \int_\beta^{\beta'} \lambda(\tilde{\beta})\mathrm{d}\tilde{\beta} - (\beta' - \beta)\lambda(\bar{\beta}) \right| \leq \frac{1}{12} \left\| \frac{d^2\lambda}{d\beta^2} \right\|_\infty |\beta' - \beta|^3. \tag{A.31}$$

By the triangle inequality combined with (A.30) and (A.31),

$$\left| r(\beta, \beta') - \int_\beta^{\beta'} \lambda(\tilde{\beta})\mathrm{d}\tilde{\beta} \right| \leq C' \|V\|_\pi^3 |\beta' - \beta|^3 + \frac{1}{12} \left\| \frac{d^2\lambda}{d\beta^2} \right\|_\infty |\beta' - \beta|^3.$$

The result follows using Lemma 22.

□

A.1.4 Corollary 5

Proof of Corollary 5. By the triangle inequality and Theorem 4,

$$\begin{aligned}
\left| \sum_{n=1}^N r(\beta_{n-1}, \beta_n) - \Lambda \right| &= \sum_{n=1}^N \left| r(\beta_{n-1}, \beta_n) - \int_{\beta_{n-1}}^{\beta_n} \lambda(\beta)\mathrm{d}\beta \right| \\
&\leq \sum_{n=1}^N C \sup_{\beta \in [\beta_{n-1}, \beta_n]} \mathbb{E}_\beta[|V|^3] |\beta_n - \beta_{n-1}|^3 \\
&\leq C \|V^3\|_\pi \|\mathcal{B}_N\|^2.
\end{aligned}$$

The last inequality used the fact that $\sum_{n=1}^N \beta_n - \beta_{n-1} = 1$. The result follows by the mean value theorem and using (2.2). □

A.1.5 Theorem 6

Proof of Theorem 6. We first note that (b) and (c) follow immediately from (a) and Corollary 2. So to prove Theorem 6 it is sufficient to show (a). By Corollary 5 we get the following estimate for

$\Lambda(\mathcal{B}_N)$,

$$\begin{aligned}\left|\Lambda(\mathcal{B}_N) - \sum_{n=1}^{N} r(\beta_{n-1},\beta_n)\right| &= \sum_{n=1}^{N} \frac{r(\beta_{n-1},\beta_n)^2}{1-r(\beta_{n-1},\beta_n)} \\ &\leq \frac{r_N^*}{1-r_N^*}\sum_{n=1}^{N} r(\beta_{n-1},\beta_n) \\ &= \frac{r_N^*}{1-r_N^*}\left(\Lambda + O\left(\frac{1}{N^2}\right)\right), \end{aligned} \tag{A.32}$$

where $r_N^* = \max_n r(\beta_{n-1},\beta_n)$ is the maximum rejection rate. Given $\mathcal{B}_N$ generated by γ, Theorem 4 implies

$$\left|r(\beta_{n-1},\beta_n) - \int_{\beta_{n-1}}^{\beta_n} \lambda(\beta)\mathrm{d}\beta\right| \leq \frac{\tilde{C}\|V^3\|_\pi}{N^3}.$$

This implies for all $n = 1,\ldots,N$ we have $r(\beta_{n-1},\beta_n)$ satisfies,

$$\begin{aligned} r(\beta_{n-1},\beta_n) &\leq \|\lambda\|_\infty|\beta_n - \beta_{n-1}| + \frac{\tilde{C}\|V^3\|_\pi\|\gamma'\|^3}{N^3}, \\ &\leq \frac{\|V\|_\pi\|\dot{\gamma}\|_\infty}{N} + \frac{\tilde{C}\|V^3\|_\pi\|\gamma'\|^3}{N^3}. \end{aligned}$$

We arrived at the last line using (2.2) and (3.13). By Taylor's theorem, $x/(1-x) = x + O(x^2)$, which implies

$$\frac{r_N^*}{1-r_N^*} \leq \frac{\|V\|_\pi\|\dot{\gamma}\|_\infty}{N} + O\left(\frac{1}{N^2}\right). \tag{A.33}$$

Finally we arrive at our estimate by combining (A.32) and (A.33). □

A.1.6 Proposition 7

Proof of Proposition 7. Let $V^{(d)}$ be the the log-likelihood ratio between $\pi_0^{(d)}$ and $\pi_1^{(d)}$. The independence structure from Equation (3.14) tells us that $V^{(d)}(x^{(d)}) = \sum_{i=1}^d V(x_i)$ where $x^{(d)} = (x_1,\ldots,x_d)$. If $X^{(d)} = (X_1,\ldots,X_d)$ and $X'^{(d)} = (X_1',\ldots,X_d')$ be drawn i.i.d. from $\pi_\beta^{(d)}$, then $\{V(X_i) - V(X_i')\}_{i=1}^d$

are i.i.d. with mean zero and variance $2\mathrm{Var}_\beta[V]$. By the central limit theorem,

$$\frac{V^{(d)}(X^{(d)}) - V^{(d)}(X'^{(d)})}{\sqrt{2d\mathrm{Var}_\beta[V]}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \frac{V(X_i) - V(X_i')}{\sqrt{2d\mathrm{Var}_\beta[V]}} \xrightarrow[d\to\infty]{} Z \sim N(0,1). \tag{A.34}$$

Thus we have

$$\begin{aligned} \lambda^{(d)}(\beta) &= \frac{1}{2}\mathbb{E}\left[\left|V^{(d)}(X^{(d)}) - V^{(d)}(X'^{(d)})\right|\right] \\ &= \frac{1}{2}\sqrt{2d\mathrm{Var}_\beta[V]}\,\mathbb{E}\left[\left|\frac{V^{(d)}(X^{(d)}) - V^{(d)}(X'^{(d)})}{\sqrt{2d\mathrm{Var}_\beta[V]}}\right|\right]. \end{aligned} \tag{A.35}$$

The sequence of variables indexed by d in the expectation in (A.35) is also uniformly integrable. This follows by noting that the second moment of the integrand in (A.35) is uniformly bounded in d:

$$\sup_d \mathbb{E}\left[\left|\frac{V^{(d)}(X^{(d)}) - V^{(d)}(X'^{(d)})}{\sqrt{2d\mathrm{Var}_\beta[V]}}\right|^2\right] = \sup_d \frac{1}{2d\mathrm{Var}_\beta[V]} \sum_{i=1}^{d} \mathrm{Var}\left[V(X_i) - V(X_i')\right] = 1.$$

By $d \to \infty$ and using (A.34) we have

$$\lim_{d\to\infty} \sqrt{\frac{2}{d\mathrm{Var}_\beta[V]}}\,\lambda^{(d)}(\beta) = \mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}, \tag{A.36}$$

To obtain the high dimensional scaling limit for Λ_d, we use Cauchy-Schwarz

$$\frac{\lambda^{(d)}(\beta)}{\sqrt{d}} = \frac{1}{2\sqrt{d}}\mathbb{E}\left[\left|V^{(d)}(X^{(d)}) - V^{(d)}(X'^{(d)})\right|\right] \le \sqrt{\frac{\mathrm{Var}_\beta[V]}{2}}. \tag{A.37}$$

Finally, (A.36), (A.37) along with dominated convergence yield

$$\lim_{d\to\infty} \frac{\Lambda^{(d)}}{\sqrt{d}} = \int_0^1 \lim_{d\to\infty} \frac{\lambda^{(d)}(\beta)}{\sqrt{d}} \mathrm{d}\beta = \frac{1}{\sqrt{\pi}} \int_0^1 \sqrt{\mathrm{Var}_\beta[V]}\,\mathrm{d}\beta.$$

□

A.2 Chapter 4

A.2.1 Theorem 10

Proof of Theorem 10. We first note that without loss of generality we can place an artificial schedule point β_n at each of the finitely many discontinuities in W_β or its first/second derivative. Thus we assume that $W_\beta(x)$ is C^2 on each interval $[\beta_{n-1}, \beta_n]$. Later in the proof it will become clear that the contributions of these artificial schedule points becomes negligible as $\|\mathcal{B}_N\| \to 0$.

Suppose we have a fixed annealing schedule $\mathcal{B}_N$, with annealing distributions $\pi_{\beta_0}, \ldots, \pi_{\beta_N}$ interpolating along the path $\pi_\beta \propto \exp(W_\beta)$. Define the spline path $\pi_\beta^N = \frac{1}{Z^N(\beta)} \exp(W_\beta^N)$ with log-likelihood W_β^N satisfying for each segment $\beta_{n-1} \leq \beta \leq \beta_n$,

$$W_\beta^N = W_{\beta_{n-1}} + \frac{\Delta W_n}{\Delta \beta_n}(\beta - \beta_{n-1}), \tag{A.38}$$

where $\Delta W_n = W_{\beta_n} - W_{\beta_{n-1}}$ and $\Delta \beta_n = \beta_n - \beta_{n-1}$. The spline path is the concatenation of the N linear paths, between $\pi_{\beta_{n-1}}$ and π_{β_n}, i.e. the log-likelihood W_β^N agrees with W_β for $\beta \in \mathcal{B}_N$, and linearly interpolates between $W_{\beta_{n-1}}$ and W_{β_n} for $\beta \in [\beta_{n-1}, \beta_n]$. The spline path approximation is important as for a fixed schedule $\mathcal{B}_N$, parallel tempering is unable to distinguish between π^N and π. Moreover, by Taylor's theorem, for all $x \in \mathcal{X}$,

$$|W_\beta^N(x) - W_\beta(x)| \leq \frac{1}{2} \sup_\beta \left| \frac{d^2 W}{d\beta^2}(x) \right| \|\mathcal{B}_N\|^2, \tag{A.39}$$

So as $N \to \infty$, (A.39) implies for all x, we have $\pi_\beta^N(x)$ converge to $\pi_\beta(x)$ uniformly in β as $\|\mathcal{B}_N\| \to 0$.

Let λ^N be the local communication barrier for π^N defined by (4.9). For $\beta \in (\beta_{n-1}, \beta_n)$, we have π_β^N is the linear path between $\pi_{\beta_{n-1}}$ and π_{β_n}. Since the global communication barrier is invariant to reparameterization, we have

$$\int_{\beta_{n-1}}^{\beta_n} \lambda^N(\beta) \mathrm{d}\beta = \Lambda_n,$$

where Λ_n is equal to the global communication barrier for the linear path between $\pi_{\beta_{n-1}}$ and π_{β_n}.

Using Theorem 4 combined and the regularity condition (4.4), we have

$$r_n = \int_{\beta_{n-1}}^{\beta_n} \lambda^N(\beta)\mathrm{d}\beta + O(\|\mathcal{B}_N\|^3). \tag{A.40}$$

By summing (A.40) from $n = 1, \ldots, N$, we have the following estimate for the cumulative rejection,

$$\sum_{n=1}^{N} r(\pi_{\beta_{n-1}}, \pi_{\beta_n}) = \Lambda(\pi^N) + O(\|\mathcal{B}_N\|^2). \tag{A.41}$$

and the round trip rate,

$$\tau(\pi, \mathcal{B}_N) = \frac{1}{2 + 2\Lambda(\pi^N)} + O(\|\mathcal{B}_N\|). \tag{A.42}$$

Lemma 24 shows $\sup_\beta |\hat{\lambda}^N(\beta) - \lambda(\beta)|$ converges uniformly to 0 as $\|\mathcal{B}_N\| \to 0$ then by dominated convergence theorem $\Lambda(\pi^N)$ converges to $\Lambda(\pi)$ uniformly as $\|\mathcal{B}_N\| \to 0$. Combining this with (A.40), (A.41), (A.42), we complete the proof by letting $\|\mathcal{B}_N\| \to 0$. □

Proof of Lemma 24

Given a regular annealing path π, for measurable function f define

$$\|f\|_\pi = \sup_\beta \mathbb{E}_\beta[|f|],$$

and given $s > 0$, define

$$C(f, s) = \|f \exp(sV_2)\|_\pi.$$

Lemma 23. *(a) Suppose there is an $\epsilon > 0$ such that $C(1, \epsilon) < \infty$, and N is large enough so that $\|\mathcal{B}_N\|^2 < \epsilon$, then*

$$\sup_{\beta \in [0,1]} \left| \frac{Z^N(\beta)}{Z(\beta)} - 1 \right| \leq C(V_2, \epsilon)\|\mathcal{B}_N\|^2. \tag{A.43}$$

and if N is large enough so that $C(V_2,\epsilon)\|\mathcal{B}_N\|^2 < 1$ also holds, then,

$$\sup_{\beta\in[0,1]}\left|\frac{Z(\beta)}{Z^N(\beta)}-1\right| \le \frac{C(V_2,\epsilon)}{1-C(V_2,\epsilon)\|\mathcal{B}_N\|^2}\|\mathcal{B}_N\|^2. \tag{A.44}$$

(b) Suppose there is an $\epsilon > 0$ such that $C(f,\epsilon) < \infty$, and if N is large enough so that $\|\mathcal{B}_N\|^2 < \epsilon$ and $C(V_2,\epsilon)\|\mathcal{B}_N\|^2 < 1$,

$$\sup_{\beta\in[0,1]}\left|\pi_\beta^N(f)-\pi_\beta(f)\right| \le \left[\frac{C(V_2,\epsilon)C(f,\epsilon)}{1-C(V_2,\epsilon)\|\mathcal{B}_N\|^2}+C(fV_2,\epsilon)\right]\|\mathcal{B}_N\|^2. \tag{A.45}$$

Proof of Lemma 23. (a) We rewrite the expression

$$\begin{aligned}\frac{Z^N(\beta)}{Z(\beta)} &= \frac{1}{Z(\beta)}\int_{\mathcal{X}}\exp\left(W_\beta^N(x)\right)\mathrm{d}x\\ &= \int_{\mathcal{X}}\exp\left(W_\beta^N(x)-W_\beta(x)\right)\pi_\beta(x)\mathrm{d}x\\ &= 1+\int_{\mathcal{X}}\left(\exp\left(W_\beta^N(x)-W_\beta(x)\right)-1\right)\pi_\beta(x)\mathrm{d}x.\end{aligned}$$

Thus using the inequality $|e^x-1| \le e^{|x|}-1$ and the spline error (A.39),

$$\begin{aligned}\left|\frac{Z^N(\beta)}{Z(\beta)}-1\right| &\le \int_{\mathcal{X}}\left(\exp\left(\left|W_\beta^N(x)-W_\beta(x)\right|\right)-1\right)\pi_\beta(x)\mathrm{d}x\\ &\le \int_{\mathcal{X}}\left(\exp\left(\|\mathcal{B}_N\|^2V_2(x)\right)-1\right)\pi_\beta(x)\mathrm{d}x\\ &= m_\beta(\|\mathcal{B}_N\|^2)-m_\beta(0),\end{aligned}$$

where $m_\beta(s) = \pi_\beta(\exp(sV_2))$ is differentiable for $0 \le s < \epsilon$, with derivative $m'_\beta(s) = \pi_\beta(V_2\exp(sV_2))$. By the mean value theorem, we have

$$\left|\frac{Z^N(\beta)}{Z(\beta)}-1\right| \le \pi_\beta(V_2\exp(\epsilon V_2))\|\mathcal{B}_N\|^2.$$

By taking supremum over $\beta\in[0,1]$, we arrive at (A.43). The bound on $|Z(\beta)/Z^N(\beta)-1|$ arises from straightforward algebraic manipulation of the above bound.

(b) We begin by rewriting $\pi_\beta^N(f)$:

$$\begin{aligned}
\pi_\beta^N(f) - \pi_\beta(f) &= \frac{1}{Z^N(\beta)} \int_{\mathcal{X}} f(x) \exp\left(W_\beta^N(x)\right) \mathrm{d}x - \pi_\beta(f) \\
&= \int_{\mathcal{X}} \left(\frac{Z(\beta)}{Z^N(\beta)} \exp\left(W_\beta^N(x) - W_\beta(x)\right) - 1 \right) f(x)\pi_\beta(x)\mathrm{d}x \\
&= \left(\frac{Z(\beta)}{Z^N(\beta)} - 1 \right) \int_{\mathcal{X}} f(x) \exp\left(W_\beta^N(x) - W_\beta(x)\right) \pi_\beta(x)\mathrm{d}x \qquad \text{(A.46)} \\
&\quad + \int_{\mathcal{X}} f(x) \left(\exp\left(W_\beta^N(x) - W_\beta(x)\right) - 1 \right) \pi_\beta(x)\mathrm{d}x. \qquad \text{(A.47)}
\end{aligned}$$

We will find bounds on each term separately. For (A.46) we use the spline error bound (A.39), and (A.44),

$$\begin{aligned}
&\left| \left(\frac{Z(\beta)}{Z^N(\beta)} - 1 \right) \int_{\mathcal{X}} f(x) \exp\left(W_\beta^N(x) - W_\beta(x)\right) \pi_\beta(x)\mathrm{d}x \right| \\
&\leq \left| \frac{Z(\beta)}{Z^N(\beta)} - 1 \right| \int_{\mathcal{X}} f(x) \exp\left(\|\mathcal{B}_N\|^2 V_2\right) \pi_\beta(x)\mathrm{d}x \\
&\leq \frac{C(V_2,\epsilon)\|\mathcal{B}_N\|^2}{1 - C(V_2,\epsilon)\|\mathcal{B}_N\|^2} C(f,\epsilon). \qquad \text{(A.48)}
\end{aligned}$$

For (A.47), we again use $|e^x - 1| \leq e^{|x|} - 1$, the spline error bound (A.39),

$$\begin{aligned}
&\int_{\mathcal{X}} f(x) \left(\exp\left(W_\beta^N(x) - W_\beta(x)\right) - 1 \right) \pi_\beta(x)\mathrm{d}x \\
&\leq \int_{\mathcal{X}} f(x) \left(\exp\left(\|\mathcal{B}_N\|^2 V_2(x)\right) - 1 \right) \pi_\beta(x)\mathrm{d}x \\
&= m_{f,\beta}(\|\mathcal{B}_N\|^2) - m_{f,\beta}(0),
\end{aligned}$$

where $m_{f,\beta}(s) = \pi_\beta(f \exp(sV_2))$ is differentiable for $0 \leq s < \epsilon$ with derivative $m'_{f,\beta}(s) = \pi_\beta(fV_2 \exp(sV_2))$. By the mean value theorem,

$$\int_{\mathcal{X}} f(x) \left(\exp\left(W_\beta^N(x) - W_\beta(x)\right) - 1 \right) \pi_\beta(x)\mathrm{d}x \leq C(fV_2,\epsilon)\|\mathcal{B}_N\|^2. \qquad \text{(A.49)}$$

Combining (A.48) and (A.49) we get (A.45).

□

Lemma 24. *If π is a regular path with local communication barrier $\lambda(\beta)$, and π^N is the spline path*

approximation with speed $\lambda^N(\beta)$*, then for all* $\epsilon > 0$ *there is a* $\delta > 0$ *such that* $\|\mathcal{B}_N\| < \delta$ *implies*

$$|\lambda^N(\beta) - \lambda(\beta)| < \epsilon.$$

Proof of Lemma 24. Adding and subtracting $\mathbb{E}\left[\left|\frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(X) - \frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(X'_\beta)\right|\right]$ within the absolute difference $2|\hat{\lambda}^N(\beta) - \lambda(t)|$ and using the triangle inequality, it can be shown that we require bounds on

$$J_{1,\beta} = \int \pi_\beta(x)\pi_\beta(y)\left|\left|\frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(y)\right| - \left|\frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(y)\right|\right| \mathrm{d}x\mathrm{d}y,$$

and

$$J_{2,\beta} = \int \left|\pi_\beta(x)\pi_\beta(y) - \pi_\beta^N(x)\pi_\beta^N(y)\right| \left|\frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_\beta^N}{\mathrm{d}\beta}(y)\right| \mathrm{d}x\mathrm{d}y.$$

For the first term, the mean value theorem implies that there exist $s, s' \in [\beta_{n-1}, \beta_n]$ (potentially functions of x and y, respectively) such that

$$J_{1,\beta} = \int \pi_t(x)\pi_t(y)\left|\left|\frac{\mathrm{d}W_s}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_{s'}}{\mathrm{d}\beta}(y)\right| - \left|\frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(y)\right|\right| \mathrm{d}x\mathrm{d}y.$$

Split the integral into the set A of $x, y \in \mathcal{X}$ where the first term in the absolute value is larger; the same analysis with the same result applies in the other case in A^c. Here, Taylor's theorem applied to $\frac{\mathrm{d}W_s}{\mathrm{d}\beta}$ about $s = \beta$ in conjunction with (4.5) implies,

$$\left|\frac{\mathrm{d}W_s}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_{s'}}{\mathrm{d}\beta}(y)\right| \leq \left|\frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_\beta}{\mathrm{d}\beta}(y)\right| + (V_2(x) + V_2(y))\|\mathcal{B}_N\|.$$

Using this and the same procedure for A^c, we have that

$$\begin{aligned} J_{1,\beta} &\leq \int \pi_t(x)\pi_t(y)(V_2(x) + V_2(y))\|\mathcal{B}_N\|\mathrm{d}x\mathrm{d}y \\ &= 2\pi_\beta(V_2)\|\mathcal{B}_N\| \\ &\leq 2\|V_2\|_\pi\|\mathcal{B}_N\|. \end{aligned}$$

This converges to 0 as $\|\mathcal{B}_N\| \to 0$.

For the second term $J_{2,\beta}$, we can again use the mean value theorem to find $s, s' \in [\beta_{n-1}, \beta_n]$ where

$$J_{2,\beta} = \int \left|\pi_\beta(x)\pi_\beta(y) - \pi_\beta^N(x)\pi_\beta^N(y)\right| \left|\frac{\mathrm{d}W_s}{\mathrm{d}\beta}(x) - \frac{\mathrm{d}W_{s'}}{\mathrm{d}\beta}(y)\right| \mathrm{d}x\mathrm{d}y,$$

and therefore via the triangle inequality, symmetry, and the $V_1(x)$ bound on the first path derivative (4.4),

$$J_{2,\beta} \leq 2\int V_1(x) \left|\pi_\beta(x)\pi_\beta(y) - \pi_\beta^N(x)\pi_\beta^N(y)\right| \mathrm{d}x\mathrm{d}y.$$

We then add and subtract $\pi_\beta(x)\pi_\beta^N(y)$ within the absolute value and use the triangle inequality again to find that

$$\begin{aligned} J_{2,\beta} &\leq 2\int \left(V_1(x) + \pi_\beta(V_1)\right) \left|\pi_\beta(x) - \tilde{\pi}_\beta(x)\right| \mathrm{d}x \\ &= 2\int \left(V_1(x) + \pi_\beta(V_1)\right) \left|1 - \frac{\pi_\beta^N(x)}{\pi_\beta(x)}\right| \pi_\beta(x)\mathrm{d}x. \end{aligned}$$

Note that by the triangle inequality and the bound $|e^x - 1| \leq e^{|x|} - 1$,

$$\left|1 - \frac{\pi_\beta^N(x)}{\pi_\beta(x)}\right| \leq \left|\frac{Z(\beta)}{Z^N(\beta)} - 1\right| \exp\left(\|\mathcal{B}_N\|^2 V_2(x)\right) + \exp(\|\mathcal{B}_N\|^2 V_2(x)) - 1.$$

Let $f = V_1 + \mathbb{E}_{\pi_t} V_1$. If N is large enough so that, $\|\mathcal{B}_N\|^2 < \epsilon$ and $C(f,\epsilon)\|\mathcal{B}_N\|^2 < 1$, then

$$J_{2,\beta} \leq 2 \sup_{s\in[0,1]} \left|\frac{Z(s)}{Z^N(s)} - 1\right| m_{f,\beta}(\|\mathcal{B}_N\|^2) + m_{f,\beta}(\|\mathcal{B}_N\|^2) - m_{f,\beta}(0),$$

where $m_{f,\beta}(s) = \pi_\beta(f\exp(sV_2))$ is differentiable for $0 \leq s < \epsilon$ with derivative $m'_{f,\beta}(s) = \pi_\beta(fV_2\exp(sV_2))$. By the mean value theorem,

$$|m_{f,\beta}(\|\mathcal{B}_N\|^2) - m_{f,\beta}(0)| \leq C(fV_2,\epsilon)\|\mathcal{B}_N\|^2.$$

Using Lemma 23 and $\sup_\beta m_{f,\beta}(\|\mathcal{B}_N\|^2) \le C(f,\epsilon)$,

$$J_{2,\beta} \le 2\left(\frac{C(V_2,\epsilon)}{1-C(V_2,\epsilon)\|\mathcal{B}_N\|^2}C(f,\epsilon) + C(fV_2,\epsilon)\right)\|\mathcal{B}_N\|^2.$$

Therefore $J_{1,\beta} + J_{2,\beta} \to 0$ as $\|\mathcal{B}_N\| \to 0$ at a rate of $O(\|\mathcal{B}_N\|)$ which completes the proof. □

A.2.2 Proposition 11

Proof of Proposition 11. Suppose π_β is a regular annealing path in $\mathcal{M}$ corresponding to a differentiable parametric curve $\eta(\beta)$, then by chain rule,

$$(\pi_\beta, \dot{\pi}_\beta) = (p_{\eta(\beta)}, \dot{\eta}(\beta)^T S_{\eta(\beta)}) = \Phi(\eta(\beta), \dot{\eta}(\beta)), \tag{A.50}$$

where $\dot{\eta}(\beta) = \frac{\mathrm{d}\eta}{\mathrm{d}\beta} \in \mathbb{R}^d$ is the velocity of the parametric curve η at β. It follows from (A.50) that Φ is onto.

To see that Φ is one-to-one, suppose $\Phi(\eta, v) = \Phi(\eta', v')$, i.e. $p_\eta = p_{\eta'}$ and $v^T S_\eta = v'^T S_{\eta'}$. Condition (R1) implies $\eta = \eta'$ and $(v - v')^T S_\eta = 0$. By taking expectations with respect to p_η and using (4.21),

$$\begin{aligned} 0 &= \mathbb{E}_\eta\left[|(v-v')^T S_\eta|^2\right] \\ &= \mathbb{E}_\eta[(v-v')^T S_\eta S_\eta^T (v-v')] \\ &= (v-v')^T I(\eta)(v-v'). \end{aligned}$$

From condition (R4) we know $I(\eta)$ is positive definite, this is only possible if $v = v'$. □

A.2.3 Proposition 12

Proof of Proposition 12. Let $\eta(\beta) = \eta + \beta v$ be a differentiable curve in Ω for a fixed $\eta \in \Omega$ and $\|v\| = 1$. Let $\pi_\beta = p_{\eta_\eta}$ be the corresponding annealing path with log-density, $W_\beta(x) = W_{\eta(\beta)}(x)$ and velocity $\dot{\pi}_\beta = v^T S_{\eta(\beta)}$. Since $\mathcal{M}$ is a regular model, we have condition (R3) implies

$$\left|\frac{\mathrm{d}W_\beta(x)}{\mathrm{d}\beta}\right| = \left|v^T \nabla W_{\eta(\beta)}(x)\right| \le V_1(x),$$

and

$$\left|\frac{d^2 W_\beta}{d\beta^2}\right| = \left|v^T \nabla^2 W_{\eta(\beta)}(x) v\right| \leq V_2(x),$$

where V_1 and V_2 satisfy (4.20) for some $\epsilon > 0$ and π_β satisfies the conditions of Theorem 10. It follows from (4.10) as $\Delta\eta = \beta v \to 0$,

$$r(p_\eta, p_{\eta+\Delta\eta}) = \lambda_r^2(\eta, \Delta\eta) + o(\|\Delta\eta\|),$$

and therefore $D = r^2$ satisfies (4.23).

It remains to show that λ_r^2 satisfies the conditions of a Finsler metric. It can be shown that regularity conditions (R3) for $\mathcal{M}$ ensure that λ satisfies (F1). Notice that for all $\eta \in \Omega$, and $x, x \in \mathcal{X}$, we have that $v \mapsto \frac{1}{2}|v^T S_\eta(x, x')|$ defines a strictly sub-additive norm for v when $S_\eta(x, x') \neq 0$. By taking expectations, with respect to η we have $\lambda_{r^2}(\eta, v)$ defined by (4.25) is a strictly sub-additive norm for v if and only if $S_\eta(x, x')$ is not identically zero. This cannot happen since (R4) ensures $\mathrm{Var}[S_\eta(X, X')] = 2I(\eta)$ is positive definite for each η, where $X, X' \sim p_\eta$. □

A.2.4 Proposition 15

Proof of Proposition 15. Since $\mathcal{M}$ is an exponential family with linearly independent sufficient statistics W_0 and W_1, we have Ω is convex subset of $\mathbb{R}^2$ with non-empty interior and the Fisher information is positive definite. It follows that $\mathcal{M}$ satisfies conditions (R1), (R2) and (R4). Since $\nabla W_\eta = W$ and $\nabla_\eta^2 W_\eta = 0$, it follows $\mathcal{M}$ is an annealing family if $V_1 = \|W\|$ satisfies condition (4.20) with $V_2 = 0$, □

A.3 Chapter 5

A.3.1 Proposition 19

Proof of Proposition 18. We will show $\mathcal{L}_W$ defines a Feller semigroup on $C([0, 1])$ by the Hille-Yosida theorem; see Kallenberg (2002, Theorem 19.11).

Indeed the first condition is satisfied since any function $f \in C([0, 1])$ can be uniformly ap-

proximated 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 $x > \delta$, for example let $\rho_\delta = \rho(x/\delta)$, $\rho(x) = g(x)/(g(x) + g(1-x))$ and $g(x) = \exp(-1/x)\mathbf{1}_{\{x>0\}}$. We choose c 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 Kallenberg (2002, Theorem 19.11) requires that for some $\mu > 0$, the set $(\mu - \mathcal{L}_W)(\mathcal{D}(\mathcal{L}_W))$ is dense in $C([0,1])$. Let $g \in C([0,1])$ be given. We apply Saranen and Seikkala (1988, Corollary 2.2), with $f(t,y) = 2\mu y - 2g$, which is clearly square integrable in t and 2μ-Lipschitz in y. Then Saranen and Seikkala (1988, Corollary 2.2) implies that for small enough $\mu > 0$ the two-point Neumann-boundary value problem

$$\mu u - \frac{1}{2}u'' = g$$
$$u'(0) = u'(1) = 0$$

admits a solution in the Sobolev space $H^2([0,1])$ of functions with square integrable first and second derivatives. This already implies that $u \in C^1([0,1])$, whereas the continuity of g and of u a priori implies the continuity of u'' since $u'' = 2\mu u - g$. Overall, for any $g \in C([0,1])$ we can find $u \in \mathcal{D}(\mathcal{L}_W)$ such that $g = (\mu - \mathcal{L}_W)g$ establishing the second condition of Kallenberg (2002, Theorem 19.11).

The third condition of Kallenberg (2002, Theorem 19.11) is that $(\mathcal{L}_W, \mathcal{D}(\mathcal{L}_W))$ satisfies the *positive maximum principle*, that is if for some $f \in \mathcal{D}(\mathcal{L}_W))$ and $x_0 \in [0,1]$ we have $f(x_0) \geq f(x) \vee 0$ for all $x \in [0,1]$, then $f''(x_0) \leq 0$. Suppose first that the maximum is attained at an interior point $x_0 \in (0,1)$; since $f \in C^2([0,1])$, by definition of $\mathcal{D}(\mathcal{L}_W))$, $f''(x_0) \geq 0$. If on the other hand the positive maximum is attained at $x_0 = 0$, suppose that $f''(0) > \epsilon$ for all $x \leq \epsilon$. Thus for $0 < y < \epsilon$ small enough, since $f'(0) = 0$ we have

$$f(y) = f(0) + \int_0^y f'(s)\mathrm{d}s = f(0) + \int_0^y \int_0^s f''(r)\mathrm{d}r\mathrm{d}y \geq f(0) + \frac{\epsilon}{2}y^2 > f(0),$$

thus arriving at a contradiction.

We have thus established that $(\mathcal{L}_W, \mathcal{D}(\mathcal{L}_W))$ satisfies all conditions of Kallenberg (2002, Theorem 19.11) and therefore generates a Feller process. □

A.3.2 Proposition 19

Proof of Proposition 19. 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 Davis (1993, 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 Lemma 22 and the fact that $\gamma \in C^1[0,1]$ we know that $\tilde{\lambda}$ is also continuous. Therefore the third condition of Davis (1993, 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

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

where the limit is uniform in θ. However by Böttcher et al. (2013, 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

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

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

$$\left|\exp\left\{-\int_0^h \tilde{\lambda}(\theta + s)\mathrm{d}s\right\} - 1 + \tilde{\lambda}(\theta)h\right| = o(h),$$

uniformly in θ, and thus

$$\begin{aligned}\frac{1}{h}f\left(\theta+h\right)\mathbb{P}^{\theta}\left[T_1\geq h\right]-f(\theta) &= \frac{1}{h}\left[f\left(\theta+h\right)\left[1-\tilde{\lambda}(\theta)h+o(h)\right]-f(\theta)\right]\\ &= \frac{1}{h}\left[f(\theta+h)-f(\theta)\right]-\tilde{\lambda}(\theta)f(\theta)+o(1).\end{aligned}$$

In addition

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

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

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

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

$$\int\mathrm{d}\theta P_U^tf(\theta)=\int_{s=0}^t\int\mathrm{d}\theta\mathcal{L}_UP_U^sf(\theta)\mathrm{d}\theta\mathrm{d}s.$$

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

$$\begin{aligned}\int\mathrm{d}\theta\mathcal{L}_Uf(\theta) &= \int_{\theta=0}^{2\pi}f'(\theta)\mathrm{d}\theta+\int_{\theta=0}^{2\pi}\tilde{\lambda}(\theta)f(Q(\theta))\mathrm{d}\theta-\int_{\theta=0}^{2\pi}\tilde{\lambda}(\theta)f(\theta)\mathrm{d}\theta\\ &= f(2\pi)-f(0)+\int_{\theta=0}^{2\pi}\tilde{\lambda}(\theta)f(Q(\theta))\mathrm{d}\theta.\end{aligned}$$

□