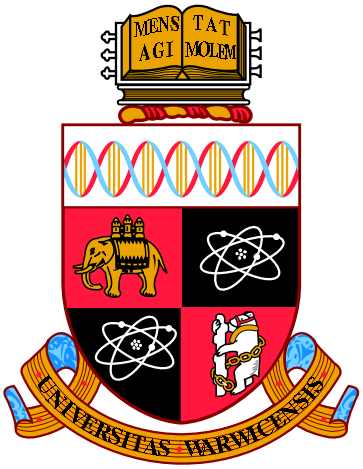

Towards Optimality of the Parallel Tempering Algorithm

by

Nicholas Tawn

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

September 2017

Contents

Acknowledgments

My supervisor, Gareth Roberts, has been integral in providing me with both excellent guidance and essential support throughout the entirety of my PhD. His inspirational input and assistance was truly appreciated.

Throughout my time at Warwick there have been many supportive friends and colleagues who have really added to my understanding and enthusiasm for statistics. There are too many to list but some key people who deserve a special mention of thanks are: Jeff Rosenthal (University of Toronto); Paul Fearnhead (University of Lancaster); Adam Johansen (University of Warwick); Murray Pollock (University of Warwick); Matt Moores (University of Warwick) and Jon Warren (University of Warwick).

Also I would like to acknowledge the financial support received from EPSRC-Engineering and Physical Sciences Research Council.

I certainly couldn't have done this without the love, care and support of both my family and fiancé Jen. Jen has been there to support, care and motivate me through the most difficult periods of the PhD process; her unwavering optimism has been fundamental to my completion of the thesis. Both my parents have been incredible with their support throughout. I am grateful that they have constantly been there to give the ultimate level of support to me throughout my life and I have needed that more than ever during the PhD. I would like them to know how grateful I am and how essential their support has been.

Declarations

I declare that the work and research contained in this thesis is my own unless otherwise stated. This thesis has been submitted to the University of Warwick only, and not to any other institution, in support of my application for the degree of Doctor of Philosophy.

Signed: _____

Nicholas Tawn

Abstract

Markov Chain Monte Carlo (MCMC) techniques for sampling from complex probability distributions have become mainstream. Big data and high model complexity demand more scalable and robust algorithms. A famous problem with MCMC is making it robust to situations when the target distribution is multi-modal. In such cases the algorithm can become trapped in a subset of the state space and fail to escape during the entirety of the run of the algorithm. This non-exploration of the state space results in highly biased sample output.

Simulated (ST) and Parallel (PT) Tempering algorithms are typically used to address multi-modality problems. These methods flatten out the target distribution using a temperature schedule. This allows the Markov chain to move freely around the state space and explore all regions of significant mass.

This thesis explores two new ideas to improve the scalability of the PT algorithm. These are implemented in prototype algorithms, QuanTA and HAT, which are accompanied by supportive theoretical optimal scaling results.

QuanTA focuses on improving transfer speed of the hot state mixing information to the target cold state. The associated scaling result for QuanTA shows that under mild conditions the QuanTA approach admits a higher order temperature spacing than the PT algorithm.

HAT focuses on preserving modal weight through the temperature schedule. This is an issue that can lead to critically poor performance of the PT approach. The associated optimal scaling result is useful from a practical perspective. The result also challenges the notion that without modal weight preservation tempering schedules can be selected based on swap acceptance rates; an idea repeatedly used in the current literature.

The new algorithms are prototype designs and have clear limitations. However, the impressive empirical performance of these new algorithms, together with supportive theory, illustrate their substantial improvement over existing methodology.

Chapter 1

Introduction

1.1 Introduction

The Bayesian approach for inference on a d-dimensional parameter x combines prior knowledge, in the form of a fully specified prior distribution, with information from observed data y, incorporated through the likelihood function, to obtain a fully specified posterior distribution of x given y, Note the unusual use of x rather than, e.g. θ, as the notation for the parameter vector; this is for the sake of consistency with all chapters in the thesis. Letting $\pi(x|y)$ denote the posterior, $f(y|x)$ the likelihood function, $\pi(x)$ the prior distribution for x and $f(y)$ the joint distribution of the data, then the posterior is computed using the standard Bayes formula:

$$\pi(x|y) = \frac{f(y|x)\pi(x)}{f(y)} \propto f(y|x)\pi(x). \tag{1.1}$$

This posterior distribution is typically intractable with the joint distribution of the data, $f(y)$, typically unknown. Tractable models can be constructed using notions of conjugacy with regards to prior distribution choice, but this defeats the object of allowing genuine expert judgment to be incorporated properly. Furthermore, the use of conjugacy is only possible in a small subset of problems.

Indeed a practitioner is typically interested in expectations of some quantity, $h(X)$, with respect to the posterior distribution, i.e.

$$\mathbb{E}_{x|y}[h(X)] = \int_{x\in\mathcal{X}} h(x)\pi(x|y)dx. \tag{1.2}$$

Due to the intractability of the posterior, computational methods are required.

Indeed this Bayesian specific issue falls into a wider class of general problems appearing across many disciplines. Specifically when a distribution $\pi(dx)$ is only

known up to a constant of proportionality but one wants to evaluate integrals of the form

$$\int_{x\in\mathcal{X}} h(x)\pi(dx).$$

A class of methods with proven success in these settings is the Monte Carlo approach. Markov Chain Monte Carlo (MCMC) is one such method that has been successfully employed in a vast range of problems. It has not only revolutionised the applicability of the Bayesian approach to statistics but also impacted subject areas across Physics, Economics, Social Sciences, Computer Science and many more. It involves the generation of a suitably constructed Markov Chain that is designed to draw samples from the intractable target distribution that is only known upto a constant of proportionality. Assuming that the algorithm has successfully sampled from the posterior distribution then estimates of moments, quantiles, ... etc can be computed by using this sample. See Section 1.2 for more details on this.

The MCMC approach assumes that the constructed Markov chain can explore the entire state space effectively in the finite run time of the chain. One major stumbling block is when the probability mass is separated into different regions in the state space. This can result in slow inter-regional exploration, or even worse, critical failure to explore all regions of significant probability mass. As a result the output sample will be biased and should not be used for Monte Carlo estimation of integrals of the form given in equation (1.2). This thesis will be primarily focused on this issue when the target distribution exhibits multi-modality.

1.1.1 Outline of the Thesis

Chapter 1 is focused on literature review and gives a basic overview of Markov chains and their application in an MCMC framework. Following this the problem of multi-modality is motivated with toy examples along with a brief overview of some of the current methods designed to overcome these issues. Detailed descriptions of the simulated and parallel tempering algorithms are given since these are the focus of development in the new work in following chapters. Chapter 1 concludes with an overview of the work in Woodard *et al.* [2009b], Woodard *et al.* [2009a] and Atchadé *et al.* [2011]. These results motivate the core novel ideas established in the following chapters.

Chapter 2 introduces a new prototype algorithm (QuanTA) designed to improve the mixing efficiency through the temperature schedule of a parallel tempering algorithm. Chapter 3 complements Chapter 2 with the development of a theoretically optimal temperature spacing result, Theorem 3.2.1, for QuanTA. Corol-

lary 3.4.1 and Theorem 3.4.1 follow from this and provide insight to the utility of QuanTA outside of the canonical Gaussian setting.

Chapter 4 introduces a new prototype algorithm (HAT) that attempts to overcome weight preservation issues prevalent when using power based tempering targets. Chapter 5 complements Chapter 4 by developing a new theoretically optimal temperature spacing result, Theorem 5.1.1, for the HAT algorithm which gives guidance to optimal temperature schedule setup. The ensuing Corollaries 5.2.1 and 5.2.2 give insight into the theorem and discuss the implications relating to the work in Atchadé *et al.* [2011], namely the major issues of using acceptance rates as a quality diagnostic.

Chapter 6 concludes the thesis with a summary of the findings followed by a discussion on two ideas for further work that naturally follow from the work in this thesis.

1.2 Markov Chains and the Metropolis-Hastings Algorithm

This section will establish the basics of the Markov chain construction and the heuristics of why such stochastic processes are so useful in a Monte Carlo framework. The basics of Markov chain theory can be found in a number of classic probability text books, e.g. Grimmett and Stirzaker [2001], Durrett [2010], but for a deep insight into Markov chain behaviour (particularly for those designing MCMC algorithms) Meyn and Tweedie [2012] is invaluable.

Heuristically, a Markov chain is a stochastic process whose evolutionary behaviour conditioned on the current value, is the same as that if it had been conditioned on the entire path history.

More formally if $(\mathcal{X}, \mathcal{B})$ is a measurable space with σ-algebra $\mathcal{B}$ then

Definition 1.2.1 (Discrete-time Markov Chain)**.** A stochastic process, X_t for $t \in \mathbb{N}$, on $\mathcal{X}$ with associated filtration $(\mathcal{F}_t)_{(t\in\mathbb{N})}$, is a discrete-time Markov chain if $\forall\ A \in \mathcal{B}$

$$\mathbb{P}(X_t \in A|\mathcal{F}_{t-1}) = \mathbb{P}(X_t \in A|\sigma(X_{t-1})).$$

Herein the associated transition kernel (for a time-homogeneous Markov chain) will be denoted

$$P(x, A) := \mathbb{P}(X_t \in A|X_{t-1} = x).$$

This construction is simplistic and tractable, hence the widespread use of this assumption in modelling applications. In many cases the key feature of interest

is the long-term behaviour of the Markov chain. Under mild conditions it turns out that the Markov chain has an ergodic behaviour and its location in the state space is described by some limiting distribution. To establish the required ergodicity results, there are three key ingredients, invariance, aperiodicity and irreducibility. For a measure μ on the space $(\mathcal{X}, \mathcal{B})$ then for any $A \in \mathcal{B}$ the following shorthand will be used herein $\mu(A) := \int_A \mu(dx)$.

Definition 1.2.2 (Invariance)**.** Suppose that $P(x, A)$ is a Markov chain transition kernel, then π is said to be invariant for the Markov chain if $\forall A \in \mathcal{B}$

$$\pi(A) = \int_{x \in \mathcal{X}} \pi(dx) P(x, A). \tag{1.3}$$

Intuitively, this says that if the location of the Markov chain at time step $t-1$ is distributed according to π then the location at time-step t is distributed according to π. It is this "π" that one hopes the chain will target in the long-run; however there is not necessarily a unique invariant distribution since the chain might get stuck/absorbed in different regions of the state space. To ensure that the chain can get everywhere repeatedly and without any cyclical behaviour the following two conditions are required on the chain:

Definition 1.2.3 (Aperiodicity and Irreducibility)**.** Recall Definition 1.2.1 of a Markov chain. Then as given in Roberts *et al.* [2004], a Markov chain with transition kernel P and invariant distribution π is

1. Aperiodic if there doesn't exist a collection of disjoint subsets of $\mathcal{X}$, $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$, of size greater than 1 such that, $\forall i \in 1, \ldots, K$ and $\forall x \in \mathcal{X}_i$ there exist $A \in \mathcal{B}$ where $A \subseteq \mathcal{X}_{\{(i+1) \bmod K\}}$ and $P(x, A) = 1$.
2. ϕ-Irreducible if there exists a non zero (σ-finite) measure, ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ there exists $n \in \mathbb{N}$ such that $P^n(x, A) > 0 \quad \forall x \in \mathcal{X}$.

With the establishment of invariance, aperiodicity and irreducibility then a key result that can be found in Roberts *et al.* [2004][Theorem 4], but originally derived in Meyn and Tweedie [2012], characterises the long-term behaviour of the Markov chain:

Theorem 1.2.1. *If a Markov chain on a state space is ϕ-irreducible, aperiodic, and has a stationary distribution π, then for π a.e. $x \in \mathcal{X}$*

$$\lim_{n \to \infty} \| P^n(x, \cdot) - \pi(\cdot) \|_{TV} = 0$$

where the norm $\|\cdot\|$ *is the TV norm defined for a measure* μ *on* $(\mathcal{X}, \mathcal{B})$ *as*

$$\| \mu(\cdot) \|_{TV} = \sup_{A \in \mathcal{B}} |\mu(A)|.$$

But why is this useful for the ultimate goal of approximating intractable integrals? For a Markov chain, X_i, with properties established in the aforementioned theorem, Meyn and Tweedie [2012] derives that for a functional $h : \mathcal{X} \to \mathbb{R}$ with $\int_{\mathcal{X}} |h(x)|\pi(dx) < \infty$ then with probability 1

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \int_{\mathcal{X}} h(x)\pi(dx) = \mathbb{E}_\pi \left[h(X)\right]. \tag{1.4}$$

A Bayesian practitioner seeks to evaluate exactly these types of integrals. Consequently, if a suitably convergent Markov chain can be established then one can use a finitely truncated version of the sample average on the LHS of equation (1.4) to estimate integrals of the form $\mathbb{E}_\pi [h(X)]$.

Setting up such a Markov chain in general could be difficult, particularly when one wants a specific limiting distribution, π, for the sample. However, having **reversibility** of the chain makes this practically possible. This is attained by selecting a transition kernel, P, such that **detailed balance** holds i.e. $\forall x, y \in \mathcal{X}$

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx). \tag{1.5}$$

It is then a routine calculation, see Roberts *et al.* [2004][Proposition 1], that the desired target π is invariant for the chain. It is exactly this setup that the famous Metropolis-Hastings algorithm utilises to achieve π-invariance.

The Metropolis-Hastings algorithm, introduced in Metropolis *et al.* [1953] and established in Hastings [1970] is arguably the most famous of the MCMC approaches.

To generate a sample of size n from the a target π, the following procedure is undertaken, with x_t denoting the value of the chain at time t:

The Metropolis-Hastings (MH) Algorithm:

- Choose an initial value for the chain, x_0.
- Choose a suitable proposal distribution for moves of the chain, denoted $q(x_{i-1}, x')$.
- Iterate over $i = 1, \ldots, n + m$

1. Propose a move $x_{i-1} \to x^{'}$ according to the proposal distribution $q(x_{i-1}, x^{'})$.
2. Compute the acceptance ratio A:

$$A := \frac{\pi(x^{'})q(x^{'}, x_{i-1})}{\pi(x_{i-1})q(x_{i-1}, x^{'})}.$$

3. Accept the move with probability

$$1 \wedge A.$$

4. If accepted set $x_i := x^{'}$ else $x_i := x_{i-1}$.

- Discard a burn-in period of the the first m samples and retain the final n.

Thus the procedure outputs a sample of points $\{x_{m+1}, \ldots, x_{n+m}\}$. Then the law of large numbers result in equation (1.4) can be exploited to estimate expectations of suitable functionals with respect to π.

1.2.1 Proposal Choice and Optimal Tuning of MH:

Clearly, due to the finite run-time of the algorithm, the choice of the proposal mechanism, $q(\cdot, \cdot)$, is fundamental in determining the sample quality.

A concrete example is for the Gaussian Random Walk Metropolis algorithm where the proposal mechanism is chosen to be a Gaussian increment centred on the current value of the chain where the user has the freedom to choose the scaling parameter, σ, that controls the variance of the proposal about the current location. There is a famous trade off, Roberts *et al.* [1997], between proposing over ambitious steps for the chain (which degrade the acceptance rates towards zero so the chain rarely moves) and being under ambitious and proposing only very small steps that have high acceptance rate but take a long time to traverse the state space (meaning very slow convergence from the initialisation point).

For a given proposal mechanism there are a number of ways that one may wish to measure the effectiveness of the resulting MH algorithm. An intuitive and principled approach, Geyer [1992], is to consider the variance of the estimator of the integral of interest, i.e. the variability of the sample estimator of $\mathbb{E}_\pi(g(x))$. The estimation of the variance of the estimator is non-trivial. For a Markov chain targeting an invariant distribution π and for some functional g the quantity $\mathrm{Var}_\pi(g(X))$ may exist (provided $g \in L^2(\pi)$) but the existence of the variance of the estimator of $\mathbb{E}_\pi(g(X))$ is not guaranteed. Denoting the t^{th} lagged auto-covariances by

$\gamma_t := \text{Cov}_\pi(g(X_0), g(X_t))$; Kipnis and Varadhan [1986] showed that for a stationary, irreducible, reversible Markov chain as $n \to \infty$

$$n\text{Var}_\pi\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) \to \sigma^2 = \sum_{t=-\infty}^{\infty} \gamma_t$$

and if this σ^2 is finite then a central limit theorem holds with

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}[g(X_i)] - \mathbb{E}_\pi(g(X))\right) \Rightarrow N(0, \sigma^2).$$

The finiteness of σ^2 is indeed an issue and there is a large amount of literature establishing when a CLT holds; even then, sample estimation of σ^2 is a difficult problem. As noted in Geyer [1992], when a spectral gap exists (see Section 1.2.2) then $\sigma^2 < \infty$.

Typically the covariances, γ_t, are positive and usually the first order covariance dominates the others in magnitude. In this case, minimising the first order covariance (by making suitably ambitious proposals for moves away from the current location) has the desirable effect of minimising the variance of the estimator. So a reasonable approach is to design/tune the proposals to maximise (on average) the distance jumped at each iteration of the chain; thereby minimising the covariance between the locations of successive chain locations. One measure of the jump/proposal ambition is the expected squared jumping distance.

Definition 1.2.4 (The Expected Squared Jumping Distance ($ESJD$))**.** For a stationary Markov chain setup to sample from an invariant distribution π, the expected squared jumping distance is given by

$$ESJD = \mathbb{E}_\pi\left[(X_1 - X_0)^2\right]$$

where X_0 and X_1 are the consecutive locations of the Markov chain under stationarity.

Sherlock [2006] uses $ESJD$ as the metric for optimal scaling of random walk Metropolis moves on spherically symmetric targets and presents a clear and insightful review on using $ESJD$ as the metric for proposal optimisation. $ESJD$ can be used to obtain theoretically optimal tuning results in simplified tractable cases. In a practical setting, empirical estimates of $ESJD$ from the output sample can be used for proposal tuning.

$ESJD$ is not always the ideal metric to judge optimal performance by. The

$ESJD$ is determining the second moment behaviour of the jump distances. There are higher order moments for the chain that could also be considered. There should be a justification for using $ESJD$ over other metrics.

As defined in Øksendal [2003], consider a diffusion process denoted by X_t with drift $\mu(X_t)$ and variance $\sigma^2(X_t)$ such that

$$dX_t = \mu(X_t)dt + \sigma^2(X_t)dB_t$$

where B_t denotes standard Brownian motion. Single components of d-dimensional Markov chains, for a number of (suitably scaled) MH algorithms, have non-trivial limiting behaviour that can be described by a diffusion process as $d \to \infty$, Roberts *et al.* [2001]. The "mixing speed" of the diffusion is determined by the volatility which is the coefficient of the Brownian motion component, i.e. $\sigma^2(X_t)$.

Roberts and Rosenthal [2014] give guidance on algorithmic tuning assuming an asymptotic diffusion can be derived. They consider two diffusion processes both with the same stationary distribution, π, but different volatility functions $\sigma_1^2(X_t)$ and $\sigma_2^2(X_t)$. They show that if $\sigma_1^2(x) \leq \sigma_2^2(x)$ (π-almost surely) then the asymptotic variance of the Monte Carlo estimator is smaller for the diffusion with volatility $\sigma_2^2(X_t)$. This suggests that if there is a limiting diffusion process for the Markov chain of interest, then optimising the volatility gives a principled approach towards minimising the estimator variability.

$ESJD$ has links with the principled limiting diffusion process approach. The $ESJD$ of a one-dimensional component of the Markov chain as the dimensionality tends to infinity converges to the quadratic variation process, i.e. $\sigma^2(X_t)$, of the limiting diffusion (if such a limit exists). Not all chains have a limiting diffusion though but this at least this gives some heuristic principal for the use of $ESJD$ over other metrics for optimisation tuning. $ESJD$ will be the metric used for optimal tuning in the theorems in both Chapters 3 and 5.

1.2.2 The Spectral Gap

Section 1.6 reviews the work of Woodard *et al.* [2009a] and Woodard *et al.* [2009b] since it is a key motivation towards the work in this thesis. A summary theorem will be given which assumes knowledge of Spectral Gap theory for Markov chain analysis. For completeness the basic definitions required to understand the summary theorem and the heuristic understanding of its implications are given.

On the space $(\mathcal{X}, \mathcal{B})$ consider a measure μ and the measure defined by the n-step transition kernel of a Markov chain (started at x) and denoted $P^n(x, \cdot)$. Herein

the following shorthand will be used:

- The measure $\mu P^n(\cdot)$ is defined as $\mu P^n(A) := \int_{\mathcal{X}} P^n(x, A)\mu(dx)\ \forall A \in \mathcal{B}$.
- For a functional f defined on $\mathcal{X}$ then $\forall x \in \mathcal{X}$, $P^n f(x) := \int_{\mathcal{X}} f(y) dP^n(x, dy)$.

Defining the following as in Woodard *et al.* [2009a]:

- The inner product (with respect to a measure π) defined on complex valued functionals on $\mathcal{X}$
$$(f, g)_\pi = \int_{\mathcal{X}} \overline{f(x)} g(x) \pi(dx). \tag{1.6}$$
- Let $L^2(\pi)$ denote the space of functionals on $\mathcal{X}$ such that $(f, f)_\pi < \infty$.
- The transition kernel, P, is defined as non-negative definite if $\forall f \in L^2(\pi)$ $(Pf, f)_\pi \geq 0$.

With these definitions in place the definition of the Spectral Gap can be introduced, as in Woodard *et al.* [2009a]:

Definition 1.2.5 (Spectral Gap)**.** For a ϕ-irreducible, aperiodic Markov chain with non-negative definite transition kernel P, invariant with respect to π, then the Spectral Gap is defined as

$$\mathbf{Gap}(P) := \inf_{f \in L^2(\pi), \mathrm{Var}_\pi(f) > 0} \left(\frac{\mathcal{E}(f, f)}{\mathrm{Var}_\pi(f)} \right) \tag{1.7}$$

where $\mathcal{E}(f, f)$ is the Dirichlet form defined as $(f, (I - P)f)_\pi$.

For the purposes of MCMC the key significance of obtaining a (bound on) the spectral gap is that it helps determine the convergence rate to invariance of the Markov chain. In fact, Woodard *et al.* [2009a] highlights that for a ϕ-irreducible, aperiodic Markov chain with non-negative definite transition kernel P, invariant with respect to π, and with initiating measure μ (and some associated constant C_μ):

$$\| \mu P^n(\cdot) - \pi(\cdot) \|_2 \leq C_\mu e^{-n\mathbf{Gap}(P)} \tag{1.8}$$

where for a measure ν which has a density with respect to π given by $\frac{d\nu}{d\pi}$ then

$$\| \nu \|_2 := \left[\int \left(\frac{d\nu}{d\pi}(x) \right)^2 \pi(dx) \right]^{1/2}.$$

Consequently, the spectral gap (if strictly positive) determines the geometric rate of convergence of the Markov chain to the invariant distribution π. The larger the value of the spectral gap the quicker the rate of convergence to invariance.

Therefore, it is desirable to maximise the spectral gap of an MCMC chain. Additionally Woodard *et al.* [2009a] and Woodard *et al.* [2009b] analyse the form of the spectral gap as dimensionality increases, thus giving an indication of how the rate of convergence decays with the curse of dimensionality.

1.2.3 A Population-Based Approach to MCMC

MCMC algorithms' performance typically rely on some tuning parameters. Examples include the variance of the proposals in the Gaussian random walk Metropolis algorithm, Roberts *et al.* [1997], or the tuning of the number of leapfrog steps and the momentum proposal in the Hamiltonian Monte Carlo algorithm.

Often, these parameters are tuned towards some theoretical optimal acceptance rate of the proposed moves or maximum empirical estimate of the $ESJD$. Traditionally, this was achieved using trial runs on test setups for the parameters that were then discarded before initiating a final run with the chosen tuned parameters. An alternative approach is to use an Adaptive MCMC framework, Roberts and Rosenthal [2007] and Roberts and Rosenthal [2009], that in a single run automates the tuning of parameters towards a value that induces a user specified chain behaviour, such as a desirable proposal acceptance rate. The complication with Adaptive MCMC is that it requires very careful implementation to ensure that the chain targets an invariant distribution that is the one desired. For complex algorithms a proof of this is non-trivial.

Adaptive MCMC approaches typically use the history of the chain (violating the Markov property). In this thesis, Chapter 2 requires the use of an approach that considers sample points under the invariant distribution other than the current location of the chain. One approach would be to use an Adaptive MCMC framework. However, to overcome the complications of an adaptive approach an alternative Population-Based approach which utilises state space augmentation was used. Such techniques have been used in Gilks *et al.* [1994], Roberts and Gilks [1994], Jasra *et al.* [2007] and involve running multiple chains in parallel, all targeting the same invariant target distribution, π, marginally whilst the proposal mechanisms for each individual chain can depend on the current location of (a subset of) the other chains. Intuitively, this can guide the direction (e.g. see the snooker algorithm of Gilks *et al.* [1994]) or shape of proposal. Importantly, the Markov property can be preserved.

For completeness, a statement and proof justifying the Population-Based

Metropolis-Hastings approach that will be used in Chapter 2 is given.

Theorem 1.2.2 (Population-Based MH Invariance)**.** *Consider a target measure $\pi(dx)$ on some measure space $(\mathcal{X}, \mathcal{B})$. On the augmented space $\mathcal{X}^n$ define the product measure*

$$\pi_n(dx_1, \ldots, dx_n) \propto \prod_{i=1}^{n} \pi(dx_i).$$

Let $A \subsetneq \{1, \ldots, n\}$ and $i \in \{1, \ldots, n\} \backslash A$. Then define a Markov chain X_n on $\mathcal{X}^n$ with component transition kernels $P_A(x_i, dy)$ taking

$$(x_1, \ldots, x_i, \ldots, x_n) \to (x_1, \ldots, y, \ldots, x_n)$$

with the associated proposal measure, defined as $q_A(x_i, dy)$, dependent on the set of locations of the chains indexed by A, i.e. $\{x_j : j \in A\}$. If the proposed moves are accepted according to the MH acceptance probability

$$\alpha_A(x_i, y) = \min \left\{ 1, \frac{\pi(y) q_A(y, x_i)}{\pi(x_i) q_A(x_i, y)} \right\}$$

then the Markov chain X_n has invariant distribution $\pi_n(dX_n)$.

Proof. The proof follows trivially as a special case of Metropolis-within Gibbs. □

1.3 The Multi-Modality Problem

Consider using the MH algorithm to target a distribution, π, that is multi-modal with the modes being well separated; Figures 1.1 and 1.2 illustrate two such examples. It is important that a chain explores the entire space so that the sample estimates are unbiased thus validating Monte Carlo integral approximations of the form (1.4). Typically the proposal mechanisms used in MH algorithms are localised and tuned to explore the local mode efficiently. However, this localisation essentially means that the Markov chain becomes trapped in a subset of the state space and (even though it theoretically satisfies all ergodicity properties in the prior section) in the finite run time it fails to explore all regions of significant probability mass. The resulting algorithm's performance in the finite run can appear similar/identical to one that fails to satisfy the irreducibility criteria introduced in Definition 1.2.3.

Figures 1.1 and 1.2 show an example where a Gaussian random walk on a toy multi-modal target for only a finite number of runs of the algorithm could result in the algorithm only exploring a restricted subset of the support of the distribution. Even though the swap acceptance rates were tuned to a theoretically suggested

optimal 0.234, Roberts *et al.* [1997], the Markov chain entirely fails to explore the full target distribution.

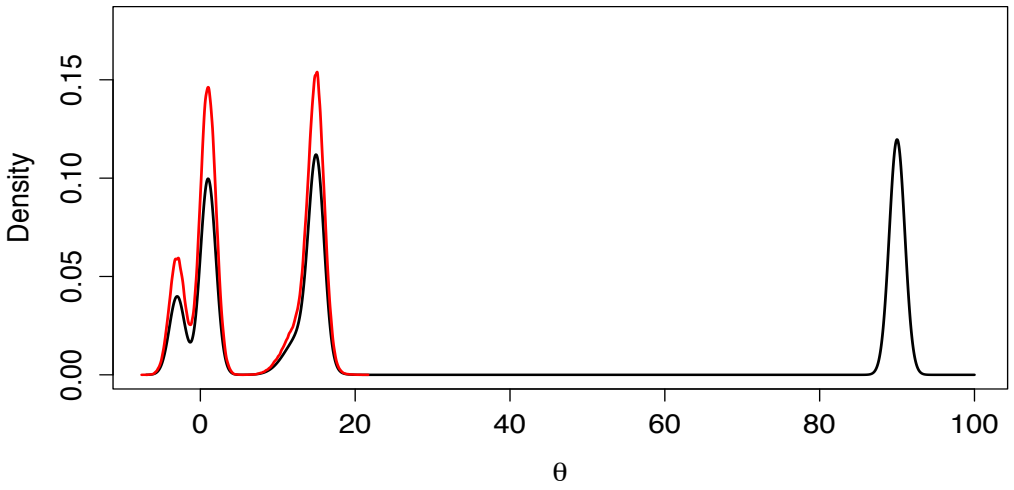

Figure 1.1: One dimensional target density (black) constructed from a Gaussian mixture distribution, over-plotted (in red) with the kernel density estimate from a 1000000 run RWM (tuned to give an approximately 0.234 acceptance rate).

Indeed many of the most effective and scalable proposal mechanisms for MH utilise the local gradient information with the Metropolis Adjusted Langevin Algorithm (MALA), e.g. Roberts and Stramer [2002], and Hamiltonian Monte Carlo (HMC), e.g. Girolami and Calderhead [2011], being arguably the most famous methods. Essentially, this class of techniques use local gradient based information that "directs" the proposed move uphill to regions of higher density. This is a heuristically sensible approach when the target is uni-modal; however, when the target distribution is multi-modal with insignificant bridging mass, then these methods tend to make the problem even worse since the gradient information helps to draw the Markov chain back towards the local mode.

In cases where the modes are "close" then approaches that incorporate ambitious proposals, e.g. Green and Mira [2001], could be used. Also, if one knows the locations of modes beforehand, then a carefully designed mixture proposal would allow for inter-modal jumps. In general, the form of the target isn't known, and if naive algorithms tuned to localised exploration are used the user will be unaware that the chain hasn't explored the entire space.

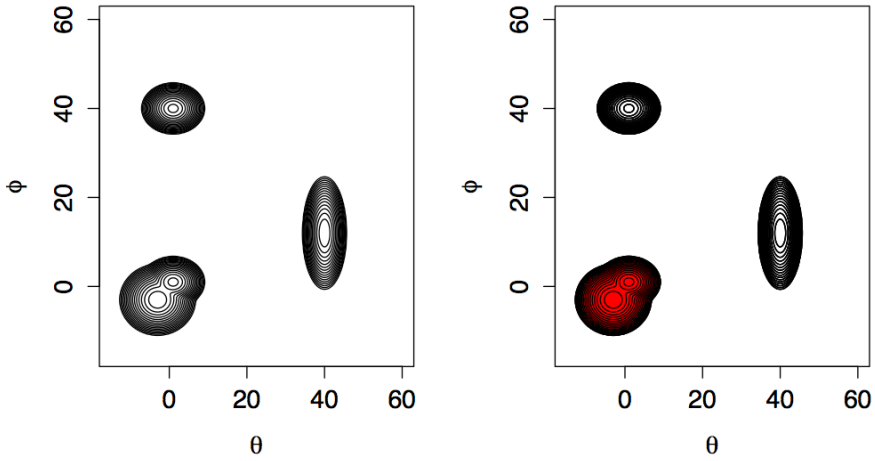

Figure 1.2: Left: 2D Gaussian mixture target density (black). Right: over-plotted (in red) with the sample of a 1000000 run RWM (tuned to give a 0.234 acceptance rate).

This motivates using a more advanced MCMC algorithm to sample from the target distribution in cases which exhibit multi-modality. The major problem is that the regions of significant probability mass do not have significant bridging mass that allows the chains to traverse along to get from region to region using localised moves. Using proposals like the Gaussian increments in RWM with large ambitious scalings in the hope the proposal will land in another modal region inevitably has extremely low acceptance rates and performance tends to deteriorate dreadfully in higher dimensions.

Definition 1.3.1 (Tempered Target at Inverse Temperature β)**.** For a target distribution $\pi(\cdot)$, the tempered target at inverse temperature β, denoted $\pi_\beta(\cdot)$, is defined as

$$\pi_\beta(x) \propto \pi(x)^\beta$$

for $\beta \in (0, \infty)$; and it is only a proper distribution provided $\int_{\mathcal{X}} \pi(x)^\beta dx < \infty$.

Such distributions are useful for a range of β values, including for optimisation problems in the simulated annealing framework where one considers $\beta \to \infty$ with an invariant Markov chain (hopefully) becoming trapped in the mode with the global maximum, Kirkpatrick *et al.* [1983]. For the purposes of sampling from a multi-modal target the range of inverse temperatures that are of interest are

$\beta \in (0, 1]$.

As $\beta \to 0$ then $\pi_\beta(\cdot)$ approaches a (potentially improper) uniform distribution on the support of the target distribution. Essentially, this has the effect of flattening out the target distribution by spreading out mass into the tails of the distribution and forming bridging mass between modes. Herein, tempered targets raised to a power $\beta \in (0, 1)$ will be will be described as hot state target distributions. This nomenclature emanates from the physics literature where, $\pi(x) \propto \exp\{-H(x)\}$ with $-H(x) < 0$ being the potential, so by heating the system to a temperature $T = 1/\beta$ (hence multiplying the potential by β) this increases the energy in the system. As such the cold state will be used herein to describe the target distribution with no heating (i.e. $\beta = 1$).

Figure 1.3 illustrates the effect of tempering on the target density used in Figure 1.1 and shows that the (normalised) tempered densities appear to approach an improper uniform distribution and have crucially provide bridging mass between the modes.

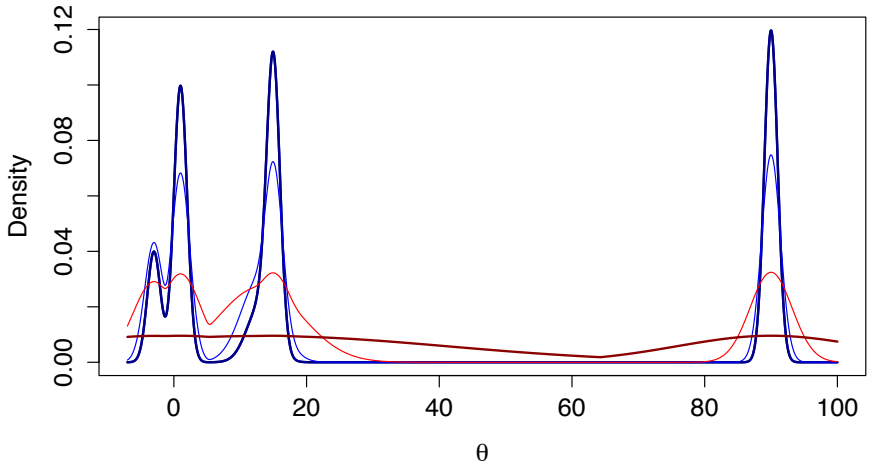

Figure 1.3: Tempered densities for β={1,0.5 ,0.1,0.005}, corresponding to (dark blue), (light blue), (red) and (dark red) respectively.

The tempered distributions have the same support and also have their modes in the same locations as the untempered distribution (since powering is a monotone function). As such, for a suitably hot temperature level target, basic methods such as RWM or HMC should be able to efficiently explore the space. Using these hot temperature targets the idea is to construct an algorithm that "borrows" information

from the mixing ability in the hotter state tempered distributions to enable full exploration of the state space in the cold level target state.

Two fundamental and widely used algorithms that exploit these tempered targets to aid mixing in multi-modal settings are the parallel and simulated tempering algorithms; these are introduced in the following Section 1.4, along with alternative and rivaling approaches for multi-modal settings.

1.4 Algorithms for Multi-Modal Target Distributions

Section 1.3 motivated the use of advanced algorithms for sampling from a multi-modal distribution. Under a given parametrisation, the target distribution is fixed and so the performance of an MCMC algorithm depends upon the proposal mechanism employed. Section 1.3 suggested that mixing information from the chains targeting the hotter, tempered distributions could be used to aid inter-modal mixing at the target cold state. Geyer [1991] and Marinari and Parisi [1992] introduced the Simulated tempering (ST) and Parallel tempering (PT) algorithms to do this. In this section the ST algorithm will be introduced and its major drawback highlighted, motivating the more practically favoured PT algorithm.

After this a brief review of alternative/competing algorithms that attempt to overcome the issues of multi-modality will be explored.

1.4.1 Simulated Tempering (ST) Algorithm

Consider a sequence of d-dimensional tempered target distributions (on a state space $\mathcal{X}$) defined at inverse temperature levels $\{\beta_0, \ldots, \beta_n\}$ where $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$. The simulated tempering approach introduced by Marinari and Parisi [1992], runs a single $(d+1)$-dimensional Markov chain, (β, X), on the extended extended state space $\{\beta_0, \ldots, \beta_n\} \times \mathcal{X}$, cycling between moves within the current temperature level and temperature level swap moves to mix through the temperature schedule. The invariant distribution of the chain, (β, X), defined on the extended state space $\{\beta_0, \ldots, \beta_n\} \times \mathcal{X}$ is

$$\pi(\beta, x) \propto K(\beta)\pi(x)^{\beta} \tag{1.9}$$

where ideally (but unrealistically) $K(\beta) = \left[\int_x \pi(x)^{\beta} \mathrm{d}x\right]^{-1}$, resulting in each temperature level being marginally normalised. The algorithm proceeds as follows:

Simulated Tempering (ST) Algorithm:

- Choose a sequence of tempering values $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$.
- Choose initial values of the chain β^0 and x^0.
- Choose the proposal mechanisms for all within temperature level type moves, denoted $q_{\beta_j}(y|x)$ for $j = 0, 1, \ldots, n$.
- Choose the number, m, of within temperature proposals the chains will perform before attempting a swap type move and choose the total number, s, of swap moves that will be attempted.
- Iterate s times:
  1. If currently in temperature level β_j uniformly randomly propose to move to one of the adjacent temperature levels, $\beta^{'} \in \{\beta_{j-1}, \beta_{j+1}\}$ say. (Denote the current position of the chain in the $\mathcal{X}$ space as x).
  2. Compute the acceptance ratio for the proposed move $(\beta_j \rightarrow \beta^{'})$ and accept the move with probability equal to
  $$\min\left(1, \frac{K(\beta^{'})\pi(x)^{\beta^{'}}}{K(\beta_j)\pi(x)^{\beta_j}}\right). \tag{1.10}$$
  3. Perform m within temperature moves to target the current inverse temperature level target using the proposal mechanism specified for the current inverse temperature value.

Choosing $K(\beta)$ to normalise the marginal temperature levels is not necessary for the running of the algorithm; indeed knowing the normalisation constants apriori is highly unlikely for most problems of interest. Figure 1.4 illustrates an example of un-normalised tempered targets when the convenient choice of $K(\beta) \propto 1$ is used. The key problem with not having marginally normalised targets is highlighted by considering the marginal distribution of the inverse temperature values. Assuming $K(\beta) \propto 1$, then by integrating out x from the joint distribution,

$$\pi(\beta) \propto \int_{\mathcal{X}} \pi(\beta, x)dx = K(\beta) \int_x \pi(x)^{\beta} \mathrm{d}x \tag{1.11}$$

and so $\pi(\beta) \propto \int_x \pi(x)^{\beta} \mathrm{d}x$. Thus the chain can end up spending dramatically different amounts of time within each temperature level; a problem exacerbated with

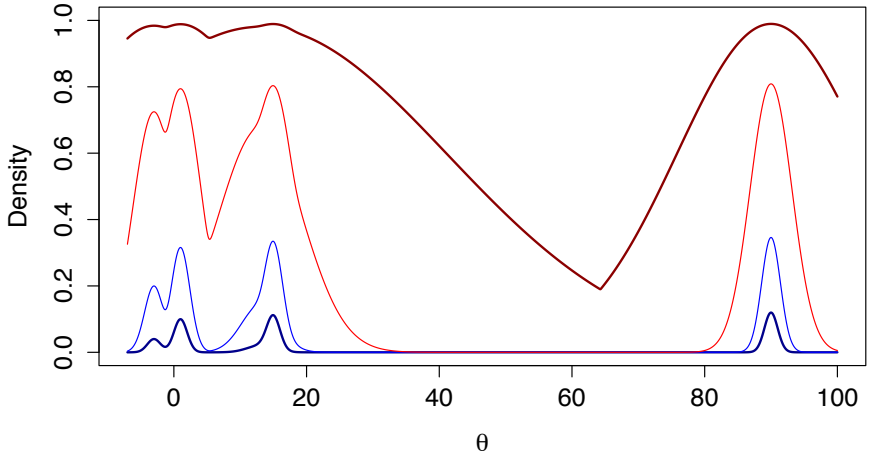

Figure 1.4: Tempered Densities for β={1,0.5,0.1,0.005}, corresponding to (dark blue), (light blue), (red) and (dark red) respectively with normalising constants $K(1) = 1$, $K(0.5) = 0.21$, $K(0.1) = 0.04$, $K(0.005) = 0.009$.

higher dimensionality. Indeed, this is an issue explored by Wang and Landau [2001] and later (and specifically for a general state space simulated tempering algorithm) Atchadé and Liu [2004]. Also, importance sampling, or more specifically, methods such as bridge sampling, Meng and Schilling [2002], can be used to attempt to approximate normalisation constants but this is a non-trivial task in multi-modal settings.

As a concrete example of the dangers of using un-normalised marginals for the simulated tempering algorithm consider Figure 1.3 with normalised tempered densities for a Gaussian mixture target distribution and the contrastingly powered up but un-normalised versions in Figure 1.4. Numerical integration allow us to compute the marginal distribution of the β's: $\pi(\beta = 0.005) = 0.77$, $\pi(\beta = 0.1) = 0.19$, $\pi(\beta = 0.5) = 0.03$, $\pi(\beta = 1) = 0.007$. Even in this single dimensional setting the Markov chain on the augmented state space would spend less that 0.7% of its time exploring the target cold state.

Example of the ST algorithm:

To demonstrate the utility of the simulated tempering algorithm, the toy problems from Figures 1.1 and 1.2 are revisited. Recall, the respective chains were unable to escape local modes when using a finite run Gaussian RWM algorithm. By using

a suitably constructed ST algorithm with a simple three level geometric schedule, $\{1, 0.05, 0.05^2\}$, (and using importance sampling to normalise the marginals), the resulting Markov chain successfully explores all modes in the target distribution as can be seen in Figure 1.5.

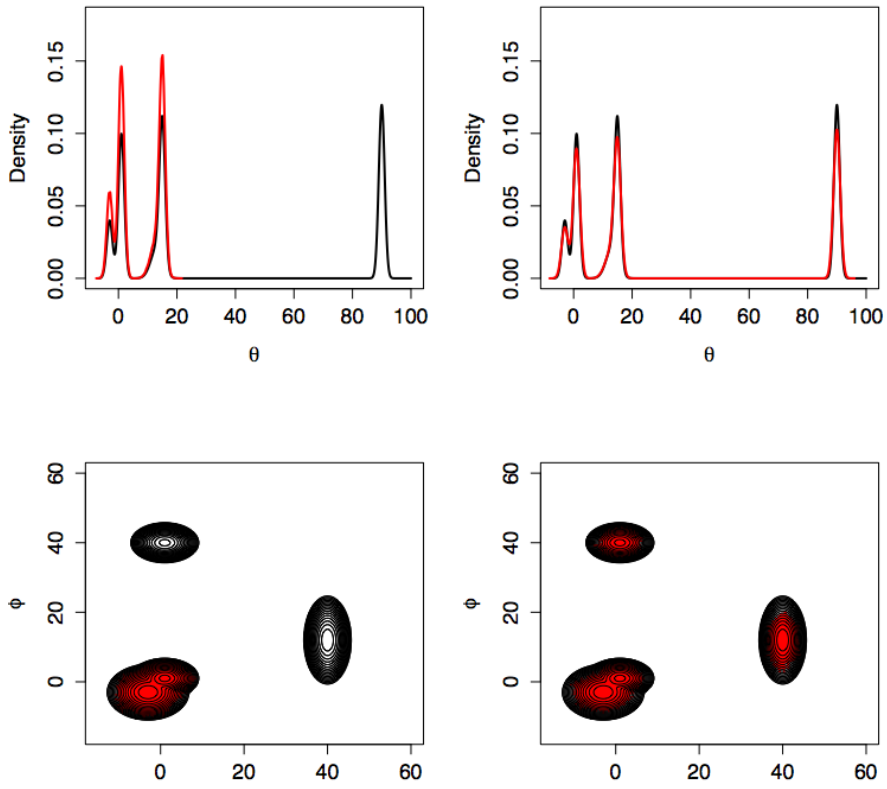

Figure 1.5: Top plots: are for the 1-dimensional example from Figure 1.1 and show the target density (black) and kernel density estimates (red) for the failure case using RWM (left) and the successful case using the ST algorithm (right). Bottom plots: are for the 2-dimensional example from Figure 1.2 with target density (black) and over plotted sample points (red), again for the failure case using RWM (left) and the successful case using the ST algorithm (right).

A convincing plot, in Figure 1.6, for the 1-dimensional example in Figure 1.5, shows that it is indeed the auxiliary hot states mixing that is allowing effective inter-modal mixing. Figure 1.6 shows the trace plot for the simulated tempering algorithm for the moves between 22000 and 24000 iterations. The background colour indicates the temperature that the chain is at, with dark red being the hottest state and blue being the cold target state. It is clear from the plot that the jumps between the

key modal areas of density in the cold state are occurring due to the mixing of the chain in the hotter states, with this mixing information then being fed back to the cold state.

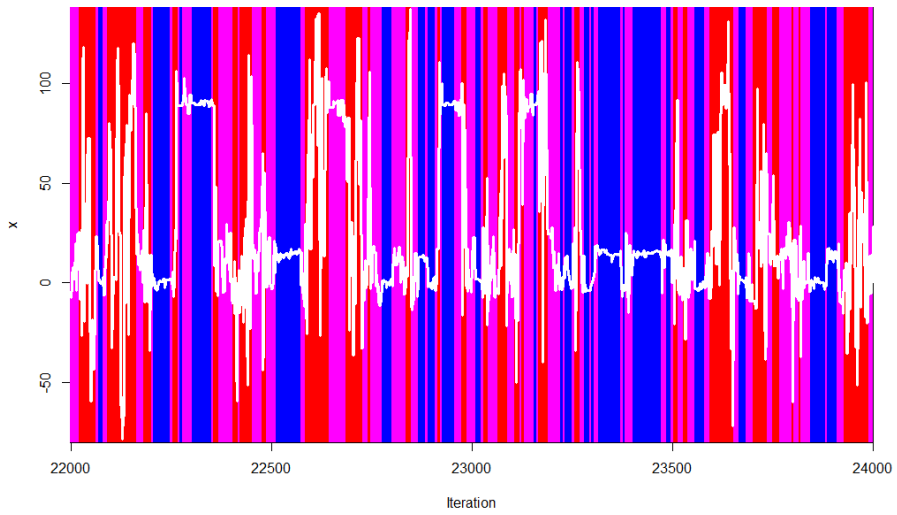

Figure 1.6: Trace plot of a simulated tempering chain targeting the 1 dimensional distribution from Figure 1.1. The inverse temperature schedule is a simple 3 level geometric schedule, $\{1, 0.05, 0.05^2\}$ and the respective temperature at each iteration is indicated by the background colour, with dark red being the hottest (0.05^2), pink the intermediate temperature (0.05) and blue the target cold state. This highlights how the chain is only able to traverse between areas of significant mass when in the hottest states.

1.4.2 Parallel Tempering (PT) Algorithm

The major practical drawback of using the ST algorithm is with regards to the lack of normalised marginals at each of the temperature levels. In toy or simple low dimensional examples numerical integration or well designed importance sampling methods could be used, although this essentially defeats the need to then use MCMC. In high dimensional, complex settings this isn't feasible.

Consider running, n, Markov chains in parallel with one at each of the different temperature levels that the previously considered ST algorithm could have used, and denote the inverse temperature levels $\{\beta_0, \ldots, \beta_n\}$ where $1 = \beta_0 > \ldots > \beta_n > 0$.

For this augmented state space, $\mathcal{X}^n$, the limiting target distribution is defined to be

$$\pi_n(x_0, x_1, \ldots, x_n) \propto \pi_{\beta_0}(x_0)\pi_{\beta_1}(x_1)\ldots\pi_{\beta_n}(x_n).$$

The chains running at the hotter temperatures have the improved fast mixing and the aim is to pass this mixing information to the cold state target chain. One approach is to propose a swap move between two chains in tandem; essentially proposing a jump in each chain to the location of the chain at the other (typically consecutive) temperature level.

Suppose a swap type move is proposed between inverse temperature levels β_j and β_k. To ensure that this swap move preserves invariance it will be accepted based on the usual Metropolis-Hastings acceptance probability,

$$\min\left(1, \frac{\pi_{\beta_j}(x_k)\pi_{\beta_k}(x_j)}{\pi_{\beta_j}(x_j)\pi_{\beta_k}(x_k)}\right). \tag{1.12}$$

Hence using this setup gives the same benefits as ST with the auxiliary hot state mixing information aiding the inter-modal mixing in the cold states. However, it is clear from equation (1.12) that the swap acceptance probability no longer depends on any marginal temperature normalisation constants since they all cancel in the ratio.

Although, very close to the ST procedure with obvious tweaking, the PT procedure is now given and it will be referred to throughout the remainder of the thesis with the shorthand notation PT:

Parallel Tempering (PT) Algorithm:

- Choose a sequence of tempering values $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$.
- Choose initial values of the chains for each temperature level, $x_0^0, x_1^0, \ldots, x_n^0$.
- Choose the proposal mechanisms for all within temperature level type moves, denoted $q_{\beta_j}(y|x)$ for $j = 0, 1, \ldots, n$.
- Choose the number, m, of within temperature proposals the chains will perform before attempting a swap type move and choose the total number, s, of swap moves that will be attempted.
- After running the chains in parallel for a burn-in period, iterate s times:

1. Uniformly randomly select a pair of adjacent temperatures, $1/\beta_j$ and $1/\beta_{j+1}$ say, for which a swap move is proposed, and where the values of the respective chains are (currently) x_j and x_{j+1}.
2. Compute the acceptance ratio for the proposed swap and accept the swap with probability equal to

$$\min\left(1, \frac{\pi(x_j)^{\beta_{j+1}}\pi(x_{j+1})^{\beta_j}}{\pi(x_j)^{\beta_j}\pi(x_{j+1})^{\beta_{j+1}}}\right). \tag{1.13}$$

3. Perform m within temperature moves for each of the $(n+1)$ chains according to their respectively specified proposal mechanisms.

Note that in this setup only swap moves between consecutive temperatures is considered. The reason for this will be apparent from the discussion on optimal tuning for the PT algorithm discussed in Section 1.5. Also note the significant opportunity to parallelise this approach so that the n chains don't require n times longer runtime; indeed, VanDerwerken and Schmidler [2013] note substantial computational gains can be made for the PT algorithm.

1.4.3 Associated and Rival MCMC Algorithms for Multi-modality

For completeness, this section discusses some of the notable and associated algorithms used in an MCMC framework that are either closely linked to the ST and PT algorithms or have rivaling approaches altogether.

- **The Equi-Energy Sampler**, Kou *et al.* [2006], Andrieu *et al.* [2007], is an approach very similar to the PT algorithm. Suppose that the target distribution has the form $\pi(x) \propto \exp\{-h(x)\}$ where $h(x)$ denotes the energy. Then suppose a sequence of energy bands are created (potentially adaptively through the run of the algorithm Schreck *et al.* [2013]) with $H_0 < H_1 \ldots < H_{n+1} = \infty$ with $H_0 < \inf_{x\in\mathcal{X}} h(x)$, hence partitioning the energy space.

  The basic idea is that multiple chains are run on a sequence of n (truncated) tempered target distributions with the i^{th} given by $\pi_i(x) \propto \exp\{-h(x) \vee H_i\}$. The samples generated at each level are grouped into the (typically predefined) energy bands partitioned by the H_i's. The process begins by only sampling from the $\pi_0(\cdot)$ distribution. After some burn-in period, sampling from $\pi_1(\cdot)$ begins in parallel but with communication with the historical samples from $\pi_0(\cdot)$ via "swap moves" similar to the PT algorithm approach. This process

continues, sequentially adding new levels until the target state is reached at $\pi_{n+1}(\cdot)$.

The "swap moves" between target levels are proposed and accepted with the same MH ratio as in the PT approach. The major difference is that the swap location from the hotter temperature level are selected uniformly from all historical locations of the hotter chain that share the same energy band indicator as the cooler chain; thus inducing high acceptance rates of the swap proposals.

- **Tempered Transitions:** Neal [1996] introduced the tempered transitions approach to sampling from a multi-modal target distribution in an MCMC framework. Like the ST and PT algorithms this method still utilises tempered target distributions. However, unlike the PT and ST algorithms, the core Markov chain requires no state-space augmentation but instead the tempered targets are interwoven into the proposal mechanism. Essentially, the proposal is made up from a sequence of proposals that create a path through the chosen sequence of tempered target levels to the hottest state and then back to the cold target state by which point the hope is that the final position is in a different mode. As will be seen, the number of temperature levels required in high dimensional tempering procedures depends on the dimension and so the proposal complexity can scale quite badly with dimension and indeed without very carefully constructed paths, the acceptance of these moves deteriorates with dimension making the method difficult to tune.

- **Mode Jumping Proposals:** Tjelmeland and Hegstad [2001] designs an algorithm that uses optimisation techniques to find a localised mode upon proposal of a large initial proposal away from the current mode; then proposing from an appropriate Gaussian distribution from the mode point. Indeed this has some interesting links with the work in Chapter 4, specifically Section 4.3.4, where optimisation methods are used. Both Al-Awadhi *et al.* [2004] and Jennison and Sharp [2006] consider similar approaches, with the former more generally for a Reversible Jump MCMC framework, and essentially use a sequence of localised MH moves to be made after an ambitious initial move has been proposed in the hope that the ensuing localised moves will drift towards the mode point. Behrens [2008] takes a slightly different approach by doing a prior scan of the statespace for modes using a series of parallel, optimising, simulated annealing procedures designed to locate modes and then using e.g mode point gradient based information, a (Gaussian) mixture model is fitted and used as

a standard proposal during the MH scheme. This has the nice feature that the actual run of the algorithm is relatively cheap but setup costs are high, and the mixture approximation's accuracy can be an issue resulting in low acceptance rates. A criticism of any scheme relying on locating modes prior to the run of the algorithm is that such approaches tend to leave no ability to adapt if modes are missed initially or if the fitted mixture proposal is poorly fitted.

- **A Repulsive-Attractive Metropolis Algorithm for Multimodality:** Tak *et al.* [2016] introduce another very interesting idea where the proposal mechanism, similar to the tempered transitions approach, is constructed from a sequence of proposals. Only localised MH moves are made but for the first stage of the sequence of moves that make up the proposal trajectory, the inverse of the target distribution π is targeted, creating a repulsion effect away from the local mode and (hopefully) then attracted towards a different mode on the latter part of the trajectory. Only a single chain is needed, but like the tempered transitions approach, it is difficult to tune the number of steps in the path of the proposal especially in high dimensional complex settings.

- Nemeth *et al.* [2017] introduce an ingenious approach that augments the state space in a way that permits a target distribution with direct bridging mass between modes. Ultimately this allows the fast mixing HMC algorithm to sample from this target by moving along the contours of the extended target density. This approach requires a computational cost which could become large since the method requires N "temperature levels" similar to the PT approach but then for the running of the HMC algorithm then augmentation to $2N$ variables is required.

- **Importance Sampling and SMC:** Additionally, and beyond the scope of this review, there's a vast range of competing approaches from the Importance Sampling and Sequential Monte Carlo (SMC) literature e.g. Neal [2001]. A nice idea that is closely linked to the PT approach is in Gramacy *et al.* [2010], where the issue of "wasted" samples in the PT algorithm arising from the augmented levels is addressed. Gramacy *et al.* [2010] proposes using these as importance samples for the cold target state and explores the optimal weighting strategy.

This thesis focuses on the development of two novel improvements to the PT algorithm. The new methodology is designed to overcome core issues that hinder the mixing speed of the chain in the PT algorithm. It is worth noting that all the

competing methods noted above suffer from the the problems noted in Chapter 4, and almost all the methods noted above suffer from the issues targeted in Chapter 2. Hence, the thesis, albeit entirely focused on the PT approach, is in many senses more broadly applicable to methodological improvement for many of the alternative multi-modal solutions.

1.5 Optimal Setup of the Temperature Schedule

In both the ST and PT algorithms, a sequence of $n+1$ inverse temperature levels were required, $\{\beta_0, \ldots, \beta_n\}$. Little was explained about why a schedule with intermediate levels, rather than simply 2 levels (a hot level and cold level) are needed. Clearly, once a sufficiently hot level has been chosen then the corresponding chain can mix through all areas of significant probability mass. Recall the acceptance probabilities for the ST and PT algorithms respectively (1.10) and (1.12); these preserve the invariance to π and if one considers the hotter state chain as providing a proposal for the next location of the colder state chain then proposals to locations that are unlikely under the target distribution are unlikely to be accepted.

As a heuristic, consider a one dimensional target distribution π which is a standard Gaussian distribution, i.e.

$$\pi(x) \propto \exp\left(-\frac{x^2}{2}\right)$$

and so at inverse temperature level β

$$\pi_\beta(x) \propto \exp\left(-\frac{\beta x^2}{2}\right)$$

hence the target distribution is still Gaussian but with variance $1/\beta$. Figure 1.7 shows the normalised density of $\pi(x)$ but is over-plotted by the (also normalised) target density at inverse temperature level $\beta = 0.1$. Clearly, locations for samples from the hotter target are very likely to occur in areas that are unrepresentative of the cold state target and hence one would expect that temperature swap move proposals from these locations would be rejected.

The bigger the gap in the temperature space between consecutive levels the larger that this issue becomes and as the gap increases the acceptance rate of swap moves between these levels diminishes to zero. Consequently, a schedule with intermediate temperatures is used to feed the mixing information from these hot states through to the cold target state.

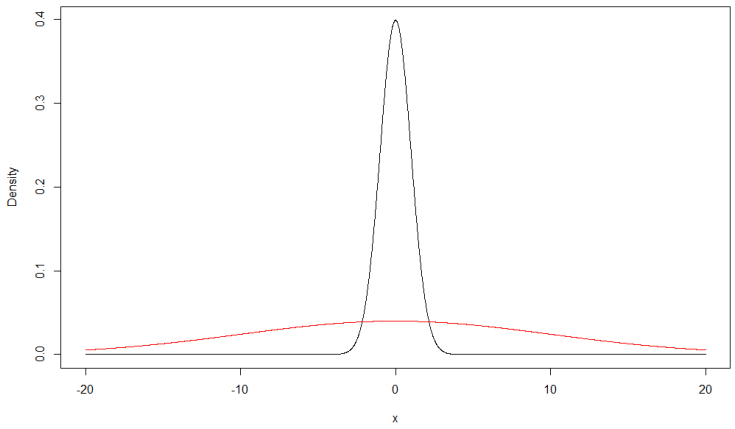

Figure 1.7: Black line: density of a standard Gaussian distribution. Red line: density of a standard Gaussian at inverse temperature level $\beta = 0.1$. Note the limited overlap of representative sample locations between the two densities.

The setup of these temperature levels is a fundamental issue to the algorithm's performance. Just like tuning the scaling of the variance when using RWM, Roberts *et al.* [1997], there is a "Goldilocks" principle.

- Making the spacings too large results in low acceptance rates and slow mixing through the temperature schedule.
- Making the spacings too small means that there are many intermediate levels to mix through on the temperature schedule to get from the hot state to the cold state; again leading to slow mixing through the temperature schedule.

This issue becomes increasingly problematic with an increase in dimensionality of the problem. To emphasise this consider a d-dimensional target distribution π which is constructed from d iid standard Gaussian marginals. The swap acceptance probability for a temperature swap move between inverse temperatures β and β' in a PT setup where $\mathbf{y} \sim \pi_{\beta'}$ and $\mathbf{x} \sim \pi_{\beta}$ is

$$\min\left\{1, \exp\left(-\frac{\beta' - \beta}{2}\sum_{i=1}^{d}\left[x_i^2 - y_i^2\right]\right)\right\} \tag{1.14}$$

but as $d \to \infty$ then by the law of large numbers $\sum_{i=1}^{d} x_i^2 \to d\ \mathbb{E}_{\beta}(X^2) = d/\beta$ and similarly $\sum_{i=1}^{d} y_i^2 \to d\ \mathbb{E}_{\beta'}(Y^2) = d/\beta'$; so for large d then (1.14) is approximately

given by

$$\min\left\{1, \exp\left(-\frac{\left(\beta' - \beta\right)^2}{2\beta\beta'}d\right)\right\}. \tag{1.15}$$

The acceptance ratio is then clearly exponentially decreasing in dimension. Consequently, for a fixed spacing $\epsilon = \beta' - \beta$, the acceptance probability of a swap between consecutive temperatures will degenerate to 0 in the limit as $d \to \infty$. If the spacing here was scaled according to the dimensionality so that $\epsilon \propto O(d^{-1/2})$, then the swap move acceptance probability in (1.15) stabilises to have a non degenerate limit as $d \to \infty$.

This dimensionality degradation and ambition of proposal trade-off motivates seeking a temperature schedule that has spacings that induce an "optimal" mixing through the temperature space.

1.5.1 Existing Optimal Scaling for Temperature Spacings Results

Atchadé *et al.* [2011] investigate the problem of selecting temperature spacings when the dimensionality, d, of the target distribution tends towards infinity. Suppose a swap move between two consecutive temperature levels, β and $\beta' = \beta + \epsilon$ is proposed. To optimise the ambition of the consecutive spacings, a heuristically sensible approach is to maximise (with respect to ϵ) the $ESJD$ (see Section 1.2 and Definition 1.2.4) through the inverse temperature space. This is the approach of Atchadé *et al.* [2011] and their specific form of $ESJD$ through the temperature schedule will be denoted herein by $ESJD_\beta$

$$ESJD_\beta = \mathbb{E}_\pi\left[(\gamma - \beta)^2\right] \tag{1.16}$$

where $\gamma = \beta + \epsilon$ for some $\epsilon > 0$ if the proposed swap is accepted and $\gamma = \beta$ otherwise. The expectation is taken with respect to π, i.e. assuming invariance has been reached. Note that $ESJD_\beta$ will be used as the target metric for optimality in both Theorems 3.2.1 and 5.1.1 later on in this thesis.

Indeed, for the PT algorithm,

$$\begin{aligned} ESJD_\beta &= \mathbb{E}_\pi\left[(\gamma - \beta)^2\right] \\ &= \epsilon^2 \times \mathbb{E}_\pi\left[\mathbb{P}(\text{Swap accepted})\right] \\ &= \epsilon^2 \times \mathbb{E}_\pi\left[\min\left(1, \frac{\pi(x_j)^{\beta_k}\pi(x_k)^{\beta_j}}{\pi(x_j)^{\beta_j}\pi(x_k)^{\beta_k}}\right)\right] =: \epsilon^2 \times ACC. \end{aligned} \tag{1.17}$$

It is worth noting for those familiar with scaling arguments that the second equality here is usually non-trivial and needs justification; in this case, with the discrete and one-dimensional nature of the temperature schedule this equality is trivial.

For tractability of optimisation, Atchadé *et al.* [2011] restrict to the set of d-dimensional target distributions with (iid) form:

$$\pi(x) \propto \prod_{i=1}^{d} f(x_i). \tag{1.18}$$

Furthermore, motivated by (1.15), for non degeneracy of the limiting behaviour of the $ESJD_\beta$ as $d \to \infty$ then the spacing, ϵ, between consecutive levels must have a form

$$\epsilon = \frac{\ell}{d^{1/2}}. \tag{1.19}$$

with ℓ a positive constant to be chosen to attain an optimal $ESJD_\beta$. Pursuit of the optimal ℓ, denoted $\hat{\ell}$, leads to Theorem 1 of Atchadé *et al.* [2011]:

Theorem 1.5.1. *For the parallel tempering algorithm, under the above setting of (1.18) and (1.19), then as $d \to \infty$ the $ESJD_\beta$ is maximised when ℓ is chosen to maximise*

$$\ell^2 \times 2\Phi\left(-\ell\sqrt{\frac{I(\beta)}{2}}\right)$$

where $I(\beta) = Var_{\pi^\beta}\big(f(x)\big)$. This optimal choice of ℓ corresponds to an acceptance rate of swap moves of 0.234 (3 d.p.). The maximised asymptotic ESJD is given by:

$$ESJD_\beta = (2/dI(\beta)) \times ACC \times [\Phi^{-1}(ACC/2)]^2. \tag{1.20}$$

Using an almost identical proof, Atchadé *et al.* [2011] showed an equivalent result for the marginally normalised ST approach, where the $ESJD_\beta$ then has the slightly different form of

$$ESJD_\beta = \epsilon^2 \times \mathbb{E}_\pi\left[\min\left(1, \frac{K(\beta+\epsilon)\pi(x)^{\beta+\epsilon}}{K(\beta)\pi(x)^\beta}\right)\right] =: \epsilon^2 \times ACC. \tag{1.21}$$

The result of which is given in Theorem 2 of Atchadé *et al.* [2011] (given here).

Theorem 1.5.2. *For the simulated tempering algorithm, under the above setting of (1.18), (1.19) and (1.21), then as $d \to \infty$ the $ESJD_\beta$ is maximised when ℓ is chosen to maximise*

$$\ell^2 \times 2\Phi\left(-\ell\frac{\sqrt{I(\beta)}}{2}\right)$$

where $I(\beta) = Var_{\pi^\beta}\big(f(x)\big)$. *This optimal choice of* ℓ *corresponds to an acceptance rate of swap moves of 0.234 (3 d.p.). The maximised asymptotic ESJD is given by:*

$$ESJD_\beta = (4/dI(\beta)) \times ACC \times [\Phi^{-1}(ACC/2)]^2. \tag{1.22}$$

Insight and Impact of Theorems 1.5.1 and 1.5.2:

Theorems 1.5.1 and 1.5.2 give explicit formulas for derivation of the optimal ℓ for consecutive temperature spacings. Albeit derived for target distributions that are assumed to have iid type construction in (1.18), for a practitioner, the associated 0.234 optimal swap acceptance rate gives powerful setup guidelines. The theorems also suggest a strategy for optimal setup starting with a chain at the hottest level and tuning the spacing to successively colder temperature levels based on the swap acceptance rate to attain consecutive swap rates close to 0.234. Indeed, using a stochastic approximation algorithm, see Robbins and Monro [1951], then Miasojedow *et al.* [2013] take an adaptive MCMC approach (see Roberts and Rosenthal [2009]) to the setup of the temperature spacings with the target being to exploit the above 0.234 tuning suggested by Atchadé *et al.* [2011].

In practice, for most interesting problems, the target will not satisfy the iid assumption (1.18). Atchadé *et al.* [2011] consider using the 0.234 rule when targeting an inhomogeneous Gaussian distribution and the Ising model, with neither example satisfying the distributional assumption in (1.18). The first of these examples illustrates that even when the distribution doesn't satisfy equation (1.18) the 0.234 rule still appears optimal when considering the empirical $ESJD_\beta$ over different spacings.

A highly insightful example is for the Ising model. Atchadé *et al.* [2011] compares the efficiency between an implementation of the 0.234 rule versus a traditionally chosen method of geometric spacing for the Ising model. The Ising model exhibits a phase transition (which is when there is a dramatic change in form of the distribution) as the temperature reaches some hot critical value. Using the traditional geometric schedule, the parallel tempering algorithm performs poorly close to this temperature. However, Atchadé *et al.* [2011] use the 0.234 rule which allocates a cluster of temperature levels around the critical temperature making it easier for the temperature swaps to "bridge" this critical temperature level (due to the temperature level move being less ambitious).

Theorems 1.5.1 and 1.5.2 give insight between the relative efficiencies of the simulated and parallel tempering schemes. The $ESJD_\beta$ for the PT algorithm given

in equation (1.20) is half that for the $ESJD_\beta$ of the ST procedure given in equation (1.21). Also, Atchadé *et al.* [2011] show that the optimal spacings between inverse temperature levels are $\sqrt{2}$ times larger for ST than for PT. This implies that compared to the PT scheme, the ST procedure is twice as efficient at mixing across the temperature space. This suggests preference towards the simulated tempering scheme. However, the optimal $ESJD_\beta$ was computed for the ST algorithm assuming the marginal normalisation constants can be found (whereas the normalising constants aren't needed for parallel tempering). Furthermore, the temperature swap in a PT scheme aids the mixing for two chains simultaneously whereas with ST mixing is only for a single chain.

Recall that in Section 1.2 the use of $ESJD$ was justified under the assumption that there is an associated limiting diffusion process. Atchadé *et al.* [2011] used the $ESJD$ but noted that full justification requires proof of existence of a limiting diffusion process. This was subsequently studied in Roberts and Rosenthal [2014]. Initially and crucially Roberts and Rosenthal [2014] establish that for two non-explosive diffusion processes X^{σ_1} and X^{σ_2} with the same invariant distribution, π, and where $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ are the variance functions, then if $\sigma_1(x) > \sigma_2(x)$ for all points x in the state space then X^{σ_1} is more efficient with respect to the asymptotic variance of estimates of $L^2(\pi)$ functionals i.e. for $f \in L^2(\pi)$

$$\lim_{T\to\infty} T^{-1/2} \operatorname{Var}\left(\int_0^T f(X_s^{\sigma_1})\, ds \right) \leq \lim_{T\to\infty} T^{-1/2} \operatorname{Var}\left(\int_0^T f(X_s^{\sigma_2})\, ds \right).$$

Roberts and Rosenthal [2014] compute the diffusion limit of the inverse temperature component, i.e. β, of an ST procedure under the same assumptions on the scaling and target form as Atchadé *et al.* [2011]. This gives a functional form of the diffusion volatility as a function of ℓ, i.e. $\sigma_\ell(\beta)$. This can be maximised with respect to ℓ at the fixed inverse temperature level β to give $\hat{\ell}(\beta)$, ensuring optimal asymptotic efficiency. Reassuringly, Roberts and Rosenthal [2014] concluded that optimal spacings are identical to those considered optimal in Atchadé *et al.* [2011] and have a corresponding expected acceptance rate of 0.234 between consecutive temperature levels.

1.5.2 The Relationship with Geometric Spacings

The problem of choosing optimal spacings has been studied previously in the physics literature e.g. Kone and Kofke [2005] and Predescu *et al.* [2004]. These papers conclude that the optimal spacings correspond to roughly a 0.23 acceptance rate

between swap moves (in agreement with the scaling theorems of Atchadé *et al.* [2011]). However, in both cases this is found for distribution functions with far more restrictive forms than that of equation (1.18).

A major observation of Atchadé *et al.* [2011] is that temperature levels should be setup consecutively. Previous studies and practitioners typically used a geometric inverse temperature schedule.

Definition 1.5.1 (Geometric Temperature Schedule)**.** A **geometric (inverse) temperature schedule** refers to a temperature schedule setup with inverse temperatures given by $1 = \beta_0 > \beta_1 > \ldots > \beta_n > 0$ where for a fixed constant $C \in (0, 1)$

$$\beta_{i+1} = C\beta_i.$$

Atchadé *et al.* [2011] show that, under their setting, the optimal temperature schedule will only be geometrically derived if $I(\beta) = \text{Var}_{f^\beta}(\log f) \propto 1/\beta^2$.

To see this, consider the optimal scaling PT result in Theorem 1.5.1 and note that the limiting $ESJD_\beta$ is given by

$$\frac{\ell^2}{d} \times 2\Phi\left(-\ell\sqrt{\frac{I(\beta)}{2}}\right).$$

Substituting in $u = \ell\sqrt{I(\beta)}$ and then maximising now with respect to u gives a value $\hat{u}$ (which importantly doesn't depend on $I(\beta)$). Then the corresponding optimal ℓ is given by

$$\hat{\ell} = \frac{\hat{u}}{\sqrt{I(\beta)}}.$$

Now for a geometric schedule then $\hat{\ell}$ must be proportional to β with the constant of proportionality not depending on the value of β. This happens if $I(\beta) \propto 1/\beta^2$. A key example is for the setting of a uni-modal iid Gaussian target since in this case $I(\beta) = 1/\beta^2$. This is the fundamental justification for the use of a geometric schedule for the canonical Gaussian empirical examples of Section 2.7, and later on, this is used in the derivation of the result in Corollary 5.2.1.

1.6 Torpid and Rapid Mixing of the PT Algorithm

Atchadé *et al.* [2011] and Roberts and Rosenthal [2014] give practical guidance to setup the temperature schedule in an optimal way. From the perspective of the ST/PT algorithm, both Atchadé *et al.* [2011] and Roberts and Rosenthal [2014], seek a Markov chain that can mix in the inverse temperature component "optimally".

Heuristically, this means that their approach seeks a setup that instigates a chain that can move from the hot state to the cold state (and vice versa) as quickly as possible. Such an approach doesn't consider how well the within temperature chains are mixing or even worse whether the temperature marginal component is only being tuned to work in a subset of the modes. The latter issue is the focus of study in Chapters 4 and 5.

The performance of the full chain for an ST/PT approach has been studied in detail in Zheng [2003], Madras and Zheng [2003], Woodard *et al.* [2009b] and Woodard *et al.* [2009a]. Essentially, these studies partition the state space into regions (typically containing a mode) and study the resulting ST Markov chain by breaking its mixing efficiency into three intuitive core components:

1. The mixing of the chain between regions at the hot state;
2. The mixing of the chain within a region;
3. The mixing of the chain through the temperature levels via the swap move.

Woodard *et al.* [2009a] and Woodard *et al.* [2009b] have the most informative results regarding the scalability of the ST and PT approaches; motivating the core strategies in this thesis. Section 1.2.2 gave the core definitions and motivation for analysing the spectral gap of a Markov chain and it is this quantity that is studied in detail in Woodard *et al.* [2009a] and Woodard *et al.* [2009b].

The spectral gap gives (a bound on) the rate of convergence of the Markov chain to invariance and so analysing its behaviour, as the dimensionality of the state space increases, indicates how robust the algorithm is to the curse of dimensionality. Inevitably, the rate of convergence will decrease as the dimensionality increases hence the spectral gap will decrease. Characterising this decrease is the focus of Woodard *et al.* [2009a] and Woodard *et al.* [2009b].

Definition 1.6.1 (Rapid and Torpid Mixing)**.** As in Woodard *et al.* [2009a] and Woodard *et al.* [2009b]:

- A Markov chain is said to be **Rapidly Mixing** if the spectral gap, defined in (1.7), decays at most polynomially quickly with respect to the state space dimensionality.
- A Markov chain is said to be **Torpidly Mixing** if the spectral gap, defined in (1.7), decays at least exponentially quickly with respect to the state space dimensionality.

Of the two types of mixing characterised in Definition 1.6.1, the preferential type of mixing is Rapid mixing which scales far less badly as the dimensionality grows with the dimensionality of the problem.

The result that is rather condemning for the scalability of the PT and ST algorithms is given in Woodard *et al.* [2009b][Corollary 3.2] and with full details to be found in their paper, it states that if the following three properties hold then the ensuing ST/PT algorithm will be Torpidly mixing. If there exists a region A and inverse temperature values $\beta^* < \beta^{**}$ such that:

1. The supremum of the conductance of A (a measure of the chain's ability to escape the local region, A) over inverse temperatures above a threshold value, β^*, is exponentially decreasing with dimension. More formally, the conductance of a set $A \in \mathcal{B}$ with respect to a target distribution measure μ is given by
$$\frac{\int_A P(x, A^c)\pi(dx)}{\pi(A)\pi(A^c)}$$
where P is the transition kernel of the Markov chain and A^c denotes the compliment of A.

2. The supremum of the persistence of A (a measure of the decrease in probability weight of A at a hot temperature relative to its weight at the cold target state) over inverse temperatures, in the range $[\beta^*, \beta^{**})$, is exponentially decreasing.

3. The supremum of the overlap (a measure of the weight indifference of A between a pair of temperature levels) for all pairings $\beta \in [0, \beta^*)$ and $\beta^{'} \in (\beta^{**}, 1]$ is exponentially decreasing with dimension.

Woodard *et al.* [2009b] illustrates that the important canonical setting (that this thesis focuses on) of the Gaussian mixture target with non-identical covariance structures is Torpidly mixing. An essential failing in this case is the persistence property, this will be a key focus of the work in Chapter 4.

Additionally, Section 2 attempts to overcome some of the issues that are problematic for the overlap property restricting the ambition of the temperature spacings meaning a less dense schedule is needed in certain settings.

It is worth noting that, Woodard *et al.* [2009a], provides an interesting result that gives conditions guaranteeing Rapid mixing for the ST and PT approaches. The quantities bounding the spectral gap from below this time are similar to those sufficient for the Torpid mixing in Woodard *et al.* [2009b]. Details are in the paper but heuristically, conditions guaranteeing Rapid mixing are: the mixing quality in

each (unimodal) region; the mixing speed of the chain at the hottest levels between regions; a variation of the persistence property, described above, decaying only polynomially with dimension; and a variation of the overlap property, described above, decaying only polynomially with dimension.

For the canonical Gaussian mixture target setting, Woodard *et al.* [2009a] illustrate that Rapid mixing can be achieved for a symmetric mode setup. This is the form of the target distributions primarily used for the empirical examples in Chapter 2 and so it is worth noting that even though the PT algorithm used in those simulations is theoretically geometrically ergodic, in practice, the performance is poor for a finite run of the algorithm.

Chapter 2

Quantile Preserved Tempering

2.1 Introduction

Papaspiliopoulos and Roberts [2003] and Papaspiliopoulos *et al.* [2007] both illustrate the importance that reparametrisation can have on the efficiency of an MCMC algorithm (especially a Gibbs sampler). Dependence structure under a particular parametrisation can mean that a Gibbs sampler finds it hard to make large moves in a particular component conditional on the value of the other parameters, Roberts and Sahu [1997]. This leads to a very poorly mixing algorithm and thus high asymptotic variances of the sample estimates. However, it is sometimes possible to reparametrise in a way that reduces (and in some cases eliminates) the dependence between parameters; thus reducing or even removing the restrictive behaviour of the component-wise Gibbs moves.

Parallel/simulated tempering algorithms can be considered similar procedures to that of a Gibbs sampler where there are deterministic/random scans which mix between the within temperature moves, in the target state space, and swap moves, through the temperature space. It will become apparent that this Gibbs style behaviour is one of the major driving factors limiting the ambitiousness of the temperature spacings leading to the concept of an optimal spacing between temperature levels, Atchadé *et al.* [2011].

The following heuristic analyses a simulated tempering procedure for a d-dimensional standard Gaussian target distribution. It illustrates exactly how a dependence between state-space location and temperature can be prohibitive to the acceptance of ambitious temperature swap moves. It is then established that a reparametrisation in this canonical case overcomes this issue entirely, allowing for arbitrarily ambitious swap moves through the temperature schedule.

Then the following sections in this chapter develop methodology to exploit this idea and make it practically useful in a multi-modal setting. Albeit derived separately and in the context of tempering, there are links with the reparametrisation techniques used in the RJMCMC context of Hastie [2005].

2.2 Gibbs Behaviour in a Tempering Setting

Consider the following example of a simulated tempering procedure on a d-dimensional Gaussian target distribution:

$$x \sim N(\mathbf{0}, I_d). \tag{2.1}$$

Hence,

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{x^T x}{2}\right), \tag{2.2}$$

and so

$$K(\beta)\pi(x)^\beta = \frac{\beta^{d/2}}{(2\pi)^{d/2}} \exp\left(-\frac{\beta x^T x}{2}\right). \tag{2.3}$$

where $K(\beta) = \int \pi^\beta(z)dz$. The tempered density is therefore still Gaussian and indeed at inverse temperature level β, $x \sim N\left(\mathbf{0}, \frac{I_d}{\beta}\right)$ and thus

$$\beta\frac{x^T x}{d} \sim \chi^2_d, \tag{2.4}$$

hence

$$\mathbb{E}\left(\frac{x^T x}{d}\right) = \frac{1}{\beta} \quad \text{and} \quad \text{Var}\left(\frac{x^T x}{d}\right) = \frac{1}{\beta^2}. \tag{2.5}$$

Suppose that the simulated tempering algorithm is currently running at the inverse temperature level β and that a temperature swap move proposal to a new level β' is made. In the standard simulated tempering algorithm, given in Section 1.4.1, this temperature swap is equivalent to the joint move

$$(\beta, x) \to (\beta', x). \tag{2.6}$$

Figure 2.1 shows the joint density function between the dimension-standardised magnitude of x, i.e. $(x^T x)/d$, and the temperature, i.e. $1/\beta$. There is a strong dependence between these two variables. Large proposals for moves in the temperature space, characterized by moves of the type given in equation (2.6) are more likely to be rejected than smaller less ambitious moves.

Suppose that instead of the move at a temperature swap being characterised

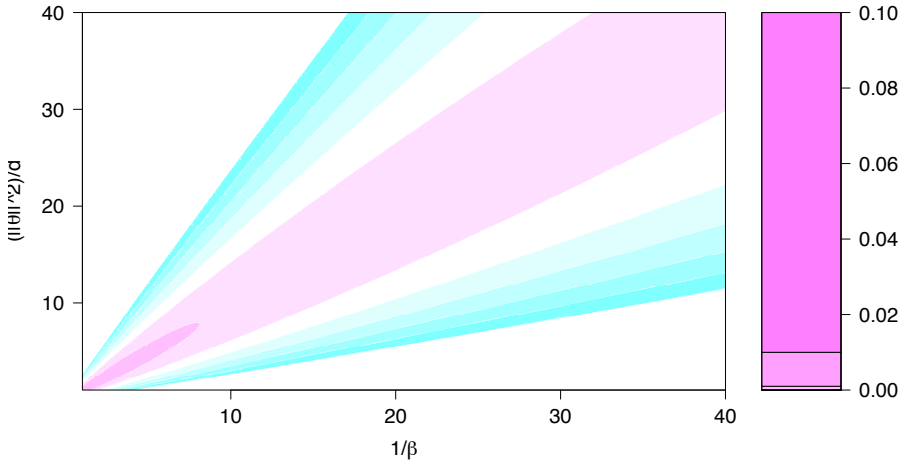

Figure 2.1: Joint distribution for $\frac{x^T x}{d}$ against $1/\beta$ when the target at the cold state is the standard normal in d-dimensions. This shows that there is a significant dependence structure between the location and temperature values as would be expected.

by equation (2.6) consider a joint move of the form

$$(\beta, x) \rightarrow (\beta^{'}, x^{'}), \tag{2.7}$$

where, conditional on the proposal of $\beta^{'}$, then the proposal for $x^{'}$ is made deterministically so that

$$x^{'} = \left(\frac{\beta}{\beta^{\prime}}\right)^{\frac{1}{2}} x. \tag{2.8}$$

The resulting acceptance rate for the swap move is then given by

$$\min\left(1, \frac{K(\beta^{'})\pi(x^{'})^{\beta^{'}}}{K(\beta)\pi(x)^{\beta}}\left|\frac{\partial x^{'}}{\partial x}\right|\right) = \min\left(1, \frac{\frac{{\beta^{'}}^{d/2}}{(2\pi)^{d/2}}\exp(-\frac{\beta^{'}{x^{'}}^{T}x^{'}}{2})}{\frac{\beta^{d/2}}{(2\pi)^{d/2}}\exp(-\frac{\beta x^T x}{2})}\left(\frac{\beta}{\beta^{\prime}}\right)^{\frac{d}{2}}\right) = 1. \tag{2.9}$$

In this example, by making the deterministic reparametrisation move for the location parameters, it is apparent that the acceptance probability of a temperature swap move is independent of the location in the state-space, $\mathcal{X}$. Atchadé *et al.* [2011] showed that in general the swap moves require spacings sizes that are $O(d^{-1/2})$ meaning that $O(d^{1/2})$ levels are required to reach a pre-specified hot state level. The reparametrisation in this case has allowed for infinitely higher order behaviour

of the scaling of the temperature spacings with respect to dimensionality; indeed, entirely overcoming the curse of dimensionality.

Figure 2.2 provides further intuition. Suppose that the current chain is at the hotter (red) state and the location of the chain is at the point indicated with the red dot. If a swap move to the colder (black) temperature is proposed and the traditional swap scheme is used, the move would very likely be rejected since the proposal is effectively trying to move a point that is "representative" in the hotter state to a location very "unrepresentative" (black point) in the colder state. What is meant by "representative" will be explained below. However, if the aforementioned reparametrisation is made in conjunction with the swap proposal then a joint move proposal is made with the location now being "representative" under the colder state (blue point).

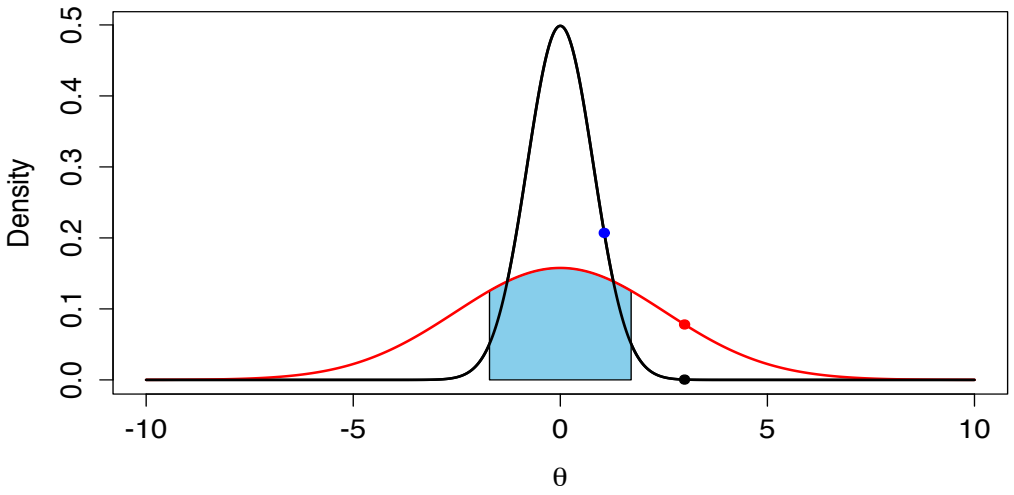

Figure 2.2: The density functions for two tempered one dimensional standard Gaussian distributions with tempering values, β, being 0.8 and 0.1, black and red respectively. The move locations of the red particle are illustrated for the standard move (black point) and the reparametrised move location (blue point).

But what does it mean for a swap move to have a "representative" location in the new temperature level? Suppose that the current location of the chain in a simulated tempering algorithm is x at inverse temperature level β and that a temperature swap move $\beta \to \beta'$ is proposed along with a reparametrised deterministic shift of $x \to x' = g(x, \beta, \beta')$. Now suppose that g is chosen in a way that pre-

serves the **quantile** between the two levels. Denoting the CDF of π^{β} by $F_{\beta}(\cdot)$ then preservation of the quantile requires g to be such that

$$F_{\beta}(x) = F_{\beta'}(g(x, \beta, \beta^{'})) \tag{2.10}$$

and so by differentiating wrt x and rearranging gives

$$1 = \frac{\pi^{\beta^{'}}(g(x, \beta, \beta^{'}))}{\pi^{\beta}(x)} \left| \frac{\partial g(x, \beta, \beta^{'})}{\partial x} \right| \tag{2.11}$$

which is exactly the acceptance ratio in the temperature swap move for the simulated tempering move. Ensuring quantile preservation gives a swap acceptance probability of 1. In this Gaussian setting one is simply making a reparametrised move that preserves the quantile value at the different levels.

Unfortunately this doesn't give a general approach to making all swap moves have acceptance probability 1 regardless of the density. For reversibility to hold one needs to ensure that the function g is a bijection with a well defined inverse. Alas, solutions to equation (2.11) are certainly not necessarily unique in any multidimensional non-trivial distribution and so it is not a good idea to solve numerically as it is unlikely to give reversible results.

In a broad class of applications it is not unreasonable to make a Gaussian approximation to a local mode. Section 2.3 will establish that this heuristic extends immediately to a target that is any general d-dimensional Gaussian. Then the rest of the chapter will explore and develop methodology to implement this simple idea to enhance the speed of the mixing through the temperature schedule in a parallel tempering setup.

2.3 A More General Reparametrisation Approach

Suppose that a more complex multi-modal distribution is now targeted. Even if the local modes can be approximated by a Gaussian distribution, it is unlikely that they will have standard covariance structure.

Suppose that the target distribution is $N(\mu, \Sigma)$ then at inverse temperature level β the hot state target will be $N(\mu, \frac{\Sigma}{\beta})$.

Consider a chain with current location, x, and suppose that a swap move $\beta \to \beta^{'}$ is proposed. The above heuristic motivates seeking a reparametrisation, $x \to x^{'}$, based on quantile preservation, to make the acceptance probability of the temperature swap independent of the location.

Defining

$$z = \left(\frac{\Sigma}{\beta}\right)^{-1/2} \left(x - \mu\right)$$

a reparametrisation of x is sought, such that, for the proposed temperature move

$$\beta \to \beta^{'} \quad x \to x^{'} \quad z \to z$$

thus keeping the quantity z constant. To this end

$$\begin{aligned} x &= \left(\frac{\Sigma}{\beta}\right)^{1/2} z + \mu \\ \text{and } \; x^{'} &= \left(\frac{\Sigma}{\beta'}\right)^{1/2} z + \mu \\ &= \left(\frac{\Sigma}{\beta'}\right)^{1/2} \left(\left(\frac{\Sigma}{\beta}\right)^{-1/2} \left(x - \mu\right)\right) + \mu \\ &= \left(\frac{\beta}{\beta'}\right)^{1/2} (x - \mu) + \mu. \end{aligned}$$

The (reparametrised) joint move is given by:

$$\beta \to \beta^{'} \quad x \to x^{'} = \left(\frac{\beta}{\beta'}\right)^{1/2} (x - \mu) + \mu. \tag{2.12}$$

A nice observation is the cancellation of the covariance matrix terms that simplify the expression to only require knowledge/approximation of the mean point μ for implementation. This type of move is similarly derived in Hastie [2005] where the aim is to adaptively fit Gaussian mixtures to different models in a RJMCMC framework and then propose swap moves between models using such moves. However, in that context the covariance structure also needs estimating since different models have entirely different covariance structures.

As already stated, to perform such a move an approximation to the local mode's mean, i.e. $\hat{\mu}$, is required. Details of how this can be done are given in Section 2.5. However, assuming there exists a technique that finds local mode points, the next section explores the utilisation of the reparametrisation idea in a parallel tempering context.

2.4 Reparametrisation for Parallel Tempering

Simulated tempering provided the heuristics but without prior knowledge or (adaptive) estimation, e.g. Atchadé and Liu [2004], of the normalisation constants of the temperature level marginals then such a procedure is impractical. Therefore, focus will be given to the parallel tempering method.

It turns out that the use of the reparametrised move in the parallel tempering setup can be very simple to implement. Essentially, there is the same setup as in the standard PT algorithm but now there is utilisation of the reparametrised move when proposing a swap move between a pair of chains at different temperature levels. All heuristics carry over from before since a PT algorithm is essentially a population with "communicating" swap move version of the simulated tempering algorithm. Consequently, moves that allow for more ambitious temperature spacings in the simulated tempering setting will allow more ambitious spacings in the PT setting.

The New Swap Move Procedure:

Suppose a swap move between two chains at temperatures β_1 and β_2 and that these are located at positions x_1 and x_2 respectively. Furthermore, assume that these locations have been assigned to local mode points μ_1 and μ_2.

Then using the reparametrisation motivated by Gaussian quantile preservation then

$$x_1 \to g(x_1, \beta_2, \beta_1, \mu_1) \qquad \text{and} \qquad x_2 \to g(x_2, \beta_1, \beta_2, \mu_2)$$

where

$$g(x, \beta', \beta, \mu) = \left(\frac{\beta}{\beta'}\right)^{1/2} (x - \mu) + \mu. \tag{2.13}$$

With d denoting the dimension of the state space, the resulting acceptance ratio required for detailed balance to hold is

$$\begin{aligned}
&\min\left(1, \frac{\pi(g(x_1, \beta_2, \beta_1, \mu_1))^{\beta_2}\pi(g(x_2, \beta_1, \beta_2, \mu_2))^{\beta_1}\left|\frac{\partial g(x_2, \beta_1, \beta_2, \mu_2)}{\partial x_1}\right|\left|\frac{\partial g(x_1, \beta_2, \beta_1, \mu_1)}{\partial x_2}\right|}{\pi(x_1)^{\beta_1}\pi(x_2)^{\beta_2}}\right) \\
&= \min\left(1, \frac{\pi(g(x_1, \beta_2, \beta_1, \mu_1))^{\beta_2}\pi(g(x_2, \beta_1, \beta_2, \mu_2))^{\beta_1}\left(\frac{\beta_2}{\beta_1}\right)^{d/2}\left(\frac{\beta_1}{\beta_2}\right)^{d/2}}{\pi(x_1)^{\beta_1}\pi(x_2)^{\beta_2}}\right) \\
&= \min\left(1, \frac{\pi(g(x_1, \beta_2, \beta_1, \mu_1))^{\beta_2}\pi(g(x_2, \beta_1, \beta_2, \mu_2))^{\beta_1}}{\pi(x_1)^{\beta_1}\pi(x_2)^{\beta_2}}\right).
\end{aligned} \tag{2.14}$$

Rather conveniently the Jacobian's determinant values (present due to the deterministic reparametrisations) cancel from the acceptance ratio.

As discussed in Section 2.3, for this move to be made feasible estimates of the means of the local modes of the particles for which a swap move is proposed are required, i.e. $\hat{\mu}$. Suggestions of how to do this are now given in Section 2.5.

2.5 The Local Mode Point Approximation

In order to perform the reparametrisation move in equation (2.13) an estimate of the mean of the local modes is required i.e. a set of locations $\{\mu_1, \ldots, \mu_K\}$ that will be used as centring points for the reparametrisation. With symmetric modes the mean and mode points coincide but in asymmetric modes then more thought is required, see Section 2.8. In fact, in a general setting where the target distribution is in C^4 and the target distribution is powered up to a super cold level, Section 3.4 justifies setting the μ_i's as the mode points of the local modes, i.e. $\frac{\partial}{\partial x}\pi(\mu_i) = 0$. This is since the spacings exhibit higher order behaviour when using the reparametrisation move at the super cold temperatures in this case.

It is an entirely non-trivial problem to obtain estimates for the μ_i's and even K, the number of modes. Fitting any mixture model adaptively as in Hastie [2005] can be quite dangerous in multi-modal settings where modes can be discovered later in the run of the algorithms. As is highlighted in Roberts and Rosenthal [2007], adaptive MCMC needs careful implementation to ensure the true distribution is being targeted and the diminishing adaptation constraints mean that if new regions of mass are discovered late on then the adaptation won't have the ability to adapt accordingly.

Section 2.6 introduces the proposed prototype algorithm, called QuanTA, which exploits the aforementioned reparametrisation. The basic idea in the setup is that through the use of a population-based setup, clustering methods exploit the information in the population to establish estimates of the mode centres, about which the reparametrisation is centred.

A population-based approach was decided upon for two major reasons. Firstly, the easily provable invariance to the true target distribution, (recall Section 1.2.3); secondly, the ability to "adapt" fully throughout the run of the algorithm upon discovery of new modal regions. Additionally, the scalability can be favorable with typically exploitable computational parallelisation. In the particular setting of interest, there are two major considerations when applying a population based scheme to estimate the mode points, μ_i:

- In order that the Markovian property of the chains is preserved it is important that only the "most recent" information is used to estimate the means.
- The estimates of the means of the local modes of the particles that are proposed for a swap move cannot depend upon the location of these particles as this would contradict the reversibility of the move.

An ideal approach would be to have a population of chains that at the point of needing the μ_i's to be estimated, one could fit a Dirichlet Process mixture model with an unspecified number of modes, e.g. Neal [2000] and Kim *et al.* [2006]. This can be done using a Gibbs sampler, switching between updates that sample a cluster number and updates for the mixture distribution parameters; this can be computationally expensive, Raykov *et al.* [2016]. Indeed, Raykov *et al.* [2016] propose a potential solution that would allow for adaptive designation of the number of modes, see the end of Section 2.5.1 for details, but this is left as further work.

2.5.1 A Weighted K-Means Clustering Approach

The work done in this thesis considers the use of the reparametrisation move in toy target distributions that are mixtures with a known number of components. Additionally, these have well separated modes. Thus, to cluster particle locations, fairly naive approaches were considered. A computationally cheap algorithm that gave good performance in these settings was sought. Hence, non-parametric, (relatively) computationally cheap clustering procedures were considered using methods from the Machine Learning literature, see Friedman *et al.* [2001].

Consequently, to find the μ_i's, N multiple versions of the parallel tempering scheme can be run in parallel and once a temperature swap type move is proposed then a K-means clustering scheme is used, e.g. Hartigan and Wong [1979]. This provides cluster centres that can then be used as either the μ_i's or as initialisation points for a localised optimisation method if the actual mode point is being sought.

There are many versions of the K-Means procedure that are implementable but the one chosen for implementation is a **Weighted K-Means Algorithm**. This method has the capacity to add weight and therefore leverage to points in the mode centres' determination. In the tempering setting this is an intuitively sensible approach. This is because one would want the particles at the colder states, where the modes are less disperse, to have more leverage in determining the mode point.

To back this up and motivate the chosen weighting strategy, consider the the following setting and its associated maximum likelihood estimator. For a set of temperatures β_j for $j = 1, \ldots, n$. If the target, π, was indeed a $N(\mu, \Sigma)$ distribution

and there are m_j draws (with the i^{th} denoted by $x_i^{\beta_j}$) from π^{β_j} for $j = 0, 1, \ldots, n$ then the maximum likelihood estimate for μ is

$$\frac{\sum_{j=1}^{n}\sum_{i=1}^{m_j} x_i^{\beta_j}\beta_j}{\sum_{j=1}^{n}\sum_{i=1}^{m_j}\beta_j}. \tag{2.15}$$

Since Gaussian approximations to the local modes are being made then it is sensible to weight the points as in the principled maximum likelihood framework.

Weighted Clustering:

The basic version of K-means groups a collection of points $x_1, \ldots, x_n$ in to M clusters $S_1, \ldots S_M$ by selecting an allocation that minimises the following objective function

$$\text{argmin}_S \left\{ \sum_{i=1}^{M}\sum_{j=1}^{n} \mathbb{1}_{\{x_j \in S_i\}} ||x_j - \mu_i||^2 \right\} \tag{2.16}$$

where μ_i is the mean of the points allocated to the set S_i, i.e.

$$\mu_i = \frac{1}{|S_i|}\sum_{j=1}^{n} x_j \mathbb{1}_{\{x_j \in S_i\}}. \tag{2.17}$$

Thus, equation (2.17) shows that all x_j's allocated to the S_j cluster have identical leverage in determining the mean, μ_j, of the j^{th} cluster.

In the population-tempering setting, with chain locations being distributed under different temperatures, this can lead to cluster centre instability. This is because the hotter state chain locations can be very far from the mode points due to the increased dispersion at these temperatures. Indeed, the Gaussian MLE formula in equation (2.15), down-weights such points in proportion to the inverse temperature value.

This motivates using a weighted version of the K-means clustering procedure with the aim being to stabilise the mean estimation of the local mode. Weighted K-means is an almost identical procedure to that of basic K-means but utilises user-defined weight allocations for points; thus designating appropriate leverage in determining the cluster centres. For the setting of interest each chain location will be allocated a weight, determined by their inverse temperature value. For a collection of n chain locations $x_1, \ldots, x_n$ at inverse temperature levels $\beta_{x_1}, \ldots, \beta_{x_n}$, the weighted K-means algorithm is as follows:

The Weighted K-Means Algorithm:

The objective function to minimise is the same as in the standard K-means procedure, given by

$$f(S) := \text{argmin}_S \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n} \mathbb{1}_{\{x_j \in S_i\}} ||x_j - \mu_i||^2 \right\}. \tag{2.18}$$

1. Choose the following: the maximum number of iterations, I; the number of centres, K; initial allocations of points (see e.g. Bradley and Fayyad [1998]), typically done by selecting K points at random to be the initial centres; and the particle weights (which for this application are the inverse temperature values, β, of the chain locations).

2. Repeat the following until either a (local) minimum for f is found or I iterations have been completed:

   i Compute for $i = 1, \ldots, K$

   $$\mu_i = \frac{\sum_{j=1}^{n} \beta_{x_j} x_j \mathbb{1}_{\{x_j \in S_i\}}}{\sum_{j=1}^{n} \beta_{x_j} \mathbb{1}_{\{x_j \in S_i\}}}. \tag{2.19}$$

   ii Re-allocate the chain locations to the new centres by allocating to the centre closest to the location in some chosen distance metric that defines the norm in the objective from equation (2.18). This gives a new allocation S.

   iii Compute $f(S)$ and check that this has reduced from the previous allocation.

3. Return centres and particle allocations.

This procedure can be implemented using the R package"FactoClass", by Elías and Del Campob [2007] which uses a modified version of the K-means algorithm of Hartigan and Wong [1979].

Clustering During the Transition Phase:

It is worth noting that during the burn-in phase of the algorithm the allocated weights can be detrimental to the performance of the algorithm. This is because the chains will not have had a chance to explore the state space properly and thus

establish themselves in the different modes. In the transitional phase it is the hotter states that expolore and discover other regions of mass and so it is these chains that should carry significant weight in the transitional phases before invariance is established.

In practice it is best that the weighted clustering is only fully incorporated after some transitional phase. In the setup of examples in this thesis this incorporation has been done linearly with respect to iteration. After a prescribed number of iterations B the algorithm will be running under a fully weighted K-means procedure as described above but for $i \in 1, \ldots, B$ the weight assigned to the j^{th} chain, x_j, at inverse temperature β_{x_j} is given by

$$\frac{1}{B}\left((B-i) + i\beta_{x_j}\right). \tag{2.20}$$

Computational Reduction:

There are many methods for initialising the centres in a K-Means algorithm, e.g. Bradley and Fayyad [1998]. Random initialisation schemes that only use the most recent chain locations ensure that the Markov property holds; this preserves invariance of the Markov Chain in the population-framework established in Section 1.2.3.

When making a temperature swap proposal using the reparametrisation move, estimates of the mode centres are required. The idea is to obtain these estimates from clustering the other chains that are running in parallel. Indeed, in the full specification of the new scheme given in the following section then multiple PT schemes will be run in parallel. It is this setup that the following ideas exploit.

For every swap proposal, clustering to discover mode centres must be done on the most recent locations of the chains to preserve the Markov property. Using all other particles that are in the other schemes for every swap proposal would lead to a very slow and computationally expensive algorithm.

Two tricks can be used to reduce the computational costs. The first one preserves invariance but the second is theoretically unjustified.

1. The parallel schemes are split into two groupings and K-means is performed on the first half to provide mean estimates for the local mode points. Using these, swap moves incorporating the reparametrisation move can be proposed for each of the schemes in the second set. The same procedure can be performed but with the reverse roles of the groups allowing swap moves for the other half. This is a more efficient way to share the mixing information between the

parallel schemes since one is not using $N-1$ schemes for clustering every time a single scheme proposes a swap type move. This setup potentially also reduces the communication costs in the case where the parallelism of the algorithm is exploited.

2. The K-means procedure is a bottleneck in computational cost and will be run regularly throughout the algorithm. It has a number of input parameters that affect its performance. As stated above, these include the initial choice of the mode centres and the number of iterations allowed for convergence. In the typical version of the K-means algorithm, the initial centres are chosen as a random selection of K of the chain locations, but in general can be selected as input parameters to the algorithm. As the number of particles increases (necessarily with problem complexity) then the procedure will require an increasing number of iterations to provide a converged algorithm, particularly in the case that the start points are initialised randomly.

   In the design of the QuanTA algorithm, introduced in the following section, the weighted K-means procedure is repeated very regularly and can potentially be a major bottleneck of the procedure in high dimensional, complex situations. However, once the population has reached invariance, it seems intuitive to pass the previous mode centres from the last clustering process to the current K-means iteration as initialisation centres. In the canonical setting of well spaced modes this can provide a major improvement in the computational overhead; with almost immediate convergence for the K-means clustering procedure.

   There is a significant catch to this though. Recycling the previous mode centres to initialise the next K-means procedure uses information from historical values of the Markov chains that are running in parallel; violating the Markov property. Empirical tests on toy examples show that this alteration reduces the computational cost without any noticeable effect on the invariant target, but much more work needs to be done to explore the validity of this.

Furthermore, albeit beyond the scope of this thesis there are a number of state of the art techniques that can improve the efficiency of the K-means procedure. For the required purposes of this thesis, the Weighted K-means procedure described above is sufficient but a practitioner should consider using more advanced techniques such as Žalik [2008], Kanungo *et al.* [2002] along with techniques that exploit parallelisation of the algorithm, see Zhao *et al.* [2009]. Indeed, Žalik [2008] gives a method that means prior selection of K is not required and uses a cost function to adaptively find K during the clustering process.

Obviously the K-means procedure can be deemed restrictive and unsuitable in many examples due to the underlying assumptions of fixed K and spherically symmetric clusters. Due to the well-spaced nature of the modes in the examples that this thesis considers then this was not an issue in the runs performed here. However, for a practitioner, an interesting and alternative approach found in Raykov *et al.* [2016] could potentially overcome these issues. It uses approximate MAP inference for Dirichlet process mixtures. The claim is that this has similar performance speed to K-means but comparable quality to Gibbs sampling. This allows fitting to clusters using a likelihood based approach which could indeed enhance the practicality of the new approach.

Having a distribution over the number of clusters would allow freedom in the determination of cluster quantity. However, it would require a carefully chosen prior over cluster numbers. Another approach would be to use a random scan through different values of K for the K-means procedure and adaptively tuning and refining which is the "best K" through an ad-hoc approach that analyses the temperature swap move acceptance rates from a particular value of K. Again this is an area of consideration for further work.

2.6 The Quantile Tempering Algorithm (QuanTA)

The setup is that there are N parallel tempering schemes running in parallel each on the same temperature levels. Herein, for this chapter, let $x^k_{(i,j)}$ represent the location of the chain on the i^{th} iteration in the k^{th} parallel scheme at the inverse temperature level β_j.

The Algorithm QuanTA:

- Choose a sequence of tempering values $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$.
- Choose initial values of the chains for each temperature level, $x^k_{00}, x^k_{01}, \ldots, x^k_{0n}$ for each k of the N parallel schemes.
- Choose the proposal mechanism for a given within temperature move, $q_{\beta_j}(x^k_{ij}, x^k_{(i+1)j})$ for $j = 1, \ldots, n$.
- Choose the value $y \in \{0, 1, \ldots, n\}$ which indexes the hottest level that the reparametrisation move will be implemented on (beyond which the standard PT swap move is used).

- Choose the number, m, of within temperature proposals the chains will perform before attempting a swap type move and choose the total number, s, of swap moves that will be attempted.

- If using a clustering procedure with a fixed cluster number, choose the number of cluster centres M.

- Iterate s times:

    1. Perform the weighted clustering procedure from Section 2.5.1 (or more generally any suitable clustering procedure) on the locations of the particles for the particles x_{ij}^k where $k \in \{1, 2, \ldots, N/2\}$ and $j \in \{1, 2, \ldots, y\}$ and i the most recent iteration of the chain. This generates cluster centres $c_1, \ldots, c_M$, from which:
        i. If the target is C^1 then use a local optimisation technique initialised at each c_j respectively to generate mode centres $\mu_1, \ldots, \mu_M$;
        ii. Else set $\mu_j = c_j \ \ \forall \ \ j = 1, \ldots, M$.
    2. For each $k \in \{N/2, \ldots, N\}$ a swap move for the k^{th} scheme is proposed. This is done as follows:
        i. Uniformly randomly select a pair of adjacent temperatures, $1/\beta_j$ and $1/\beta_{j+1}$ say, for which a reparametrised swap move will be proposed, and where the values of the respective chains are (currently) x_{ij}^k and $x_{i(j+1)}^k$. If the tempering level is too high i.e. $(j+1) > y$ then propose the standard swap move and accept with the ratio given in equation (2.36). Otherwise continue with the reparametrisation move proposal.
        ii. Classify the clusters to which the particles x_{ij}^k and $x_{i(j+1)}^k$ belong, denoted by μ_1 and μ_2 respectively.
        iii. The transformed locations are computed as in equation (2.13) for the reparametrisation move i.e.

$$g(x_{ij}^k, \beta_{(j+1)}, \beta_j, \mu_1) \text{ and } g(x_{i(j+1)}^k, \beta_j, \beta_{(j+1)}, \mu_2).$$

        iv. Compute the acceptance ratio for the proposed swap and accept the swap with probability equal to

$$\min\left(1, \frac{\pi(g(x_{ij}^k, \beta_{(j+1)}, \beta_j, \mu_1))^{\beta_{(j+1)}} \pi(g(x_{i(j+1)}^k, \beta_j, \beta_{(j+1)}, \mu_2))^{\beta_j}}{\pi(x_1)^{\beta_1} \pi(x_2)^{\beta_2}}\right).$$

3. Now perform the chosen clustering procedure on the locations of the particles for the particles x_{ij}^k where $k \in \{N/2, \ldots, N\}$ and $j \in \{1, 2, \ldots, y\}$ and i the most recent iteration of the chain.
4. Repeat the procedure in step 2 but now for $k \in \{1, 2, \ldots, N/2\}$.
5. For each of the N parallel schemes, perform m within temperature moves for each of the $(n+1)$ chains according to the proposal mechanism specified.

Note that the QuanTA algorithm described is suitable for parallelisation; both for the within temperature moves which do not require "communication" between the particles and also parallelisation between the groupings so that one group performs the within moves while the other group performs the clustering operation.

2.7 Examples of Implementation

Motivated by the heuristic provided in Section 2.2 it would be hoped that this new algorithm leads to high (close to 1) acceptance rates for swap moves when the target is a mixture of symmetric Gaussian modes.

Section 1.5 explains that even under an optimal setup the traditional PT scheme becomes increasingly expensive as dimensionality grows. As the dimensionality, d, grows the optimal spacing of the PT algorithm decays as $O(d^{-1/2})$. This means that the time taken to pass the information from the hotter mixing states to the coldest state is $O(d)$ due to the random walk nature of the swap moves, Roberts and Rosenthal [2014]. It will be shown in Chapter 3, in particular Theorem 3.2.1 and Theorem 3.4.1, that this new scheme is still $O(d)$ with spacings $O(d^{-1/2})$ for a general mode structure. In addition to this, if the target is C^4, then at colder temperatures when the Gaussian approximation to the mode becomes increasingly accurate, e.g. Barndorff-Nielsen and Nielsen [1989] and Olver [1968], then this method has a higher order behaviour in spacings with respect to the inverse temperature value, see Section 3.4.1.

The QuanTA reparametrisation move doesn't solve all the issues inherent in the PT framework. This will be highlighted with the final example in this section. In fact, Woodard *et al.* [2009b] shows that for most "interesting" examples the mixing decays exponentially slowly in dimension and the reader is directed to Chapter 4 for heuristics underlying this issue and a prototype attempt to overcome this challenging problem.

Basic Setups for the Examples in this Section:

For clarity of the new QuanTA algorithm's gains the main two examples will focus on target distributions with symmetric (i.e. all modes have the same covariance structure) Gaussian modes all with equal weights. Subsequently there will be an example where the target distribution is still an evenly weighted mixture of Gaussians but now the inter-modal variances differ between modes. This example will be a motivating example to the ideas explored in Chapter 4.

In each of the examples given, both the new QuanTA and standard (PT) parallel schemes will be run for comparison of performance. Recall the standard (PT) parallel scheme refers to the algorithm detailed in Section 1.4.2. In all examples:

1. Both the new QuanTA and PT versions were run 10 times to ensure replicability.

2. Both versions were run with the same within to swap move ratio; in all examples the algorithms performed 3 within move proposals to every 1 swap move proposal.

3. Both versions use the same set of (geometrically generated) temperature spacings; chosen to be overly ambitious for the PT setup but demonstrably underambitious for the new QuanTA scheme.

4. In addition to the overambitious schedule for the PT approach, the optimal temperature spacing for the PT setup is presented. This is to highlight the extra number of levels (and hence computational cost) needed for the PT approach to be effective in the examples. This has been found using repeated runs of the PT algorithm with temperatures selected so that there is a fixed hottest level and then levels are added (or discarded) until the suggested optimal acceptance rate of 0.234 for the PT algorithm, Atchadé *et al.* [2011], is reached.

5. For all runs the within level temperature proposals were made with Gaussian RWM moves and at each level the respective acceptance rates were tuned approximately towards the suggested optimal 0.234.

2.7.1 One-Dimensional Example

Target distribution given by:

$$\pi(x) \propto \sum_{k=1}^{5} w_k \phi_{(\mu_k, \sigma^2)}(x) \tag{2.21}$$

where $\phi_{(\mu,\sigma^2)}(.)$ is the density function of a univariate Gaussian with mean μ and variance σ^2. In this example, $\sigma = 0.01$, the mode centres are given by $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-200, -100, 0, 100, 200)$ and all modes are equally weighted with $w_1 = w_2 = \ldots = w_5$.

Hence, with very narrow well spaced modes this is a hard example for the PT algorithm but essentially canonical for QuanTA.

The temperature schedule for this example is derived from a geometric schedule (see Section 1.5.2) with an ambitious 0.0002 common ratio for the spacings. Only 3 levels are used and so the temperature schedule is given by $\{1, 0.0002, 0.0002^2\}$. Figure 2.3 gives the plot of the non-normalised target distributions at each of these 3 levels.

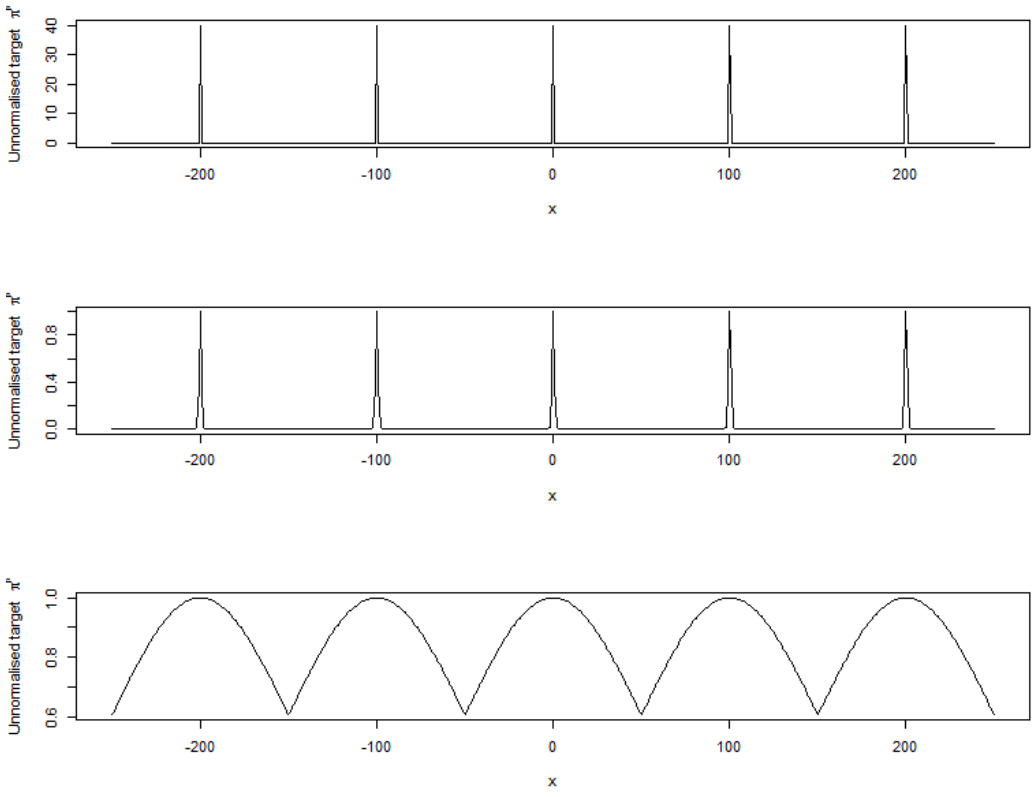

Figure 2.3: From top to bottom, plots of the target distribution given in equation (2.21) at each of the tempered levels with inverse temperatures $\{1, 0.0002, 0.0002^2\}$ respectively.

Both the PT and QuanTA algorithms were run so that 20,000 swap moves

would be attempted. For QuanTA this would be 20,000 swaps for each of the N individual parallel tempering schemes in parallel of which there were $N = 100$ in this example. Hence for a **single** scheme in each setting there is the same frequency and quantity of swap proposals. Furthermore, in all runs **all** the chains were started from a start location of -200. This bias favours the PT algorithm since it makes it hard for the clustering procedure to establish and branch out early on.

Specific to the setup of the QuanTA scheme:

- In the clustering steps of the algorithm then **eligible chains at all temperature levels** are used in the weighted clustering procedure, i.e. there is no cutoff beyond which hotter temperature chains are not considered.
- All swap moves between the levels used the reparametrisation move (conditional on being reversible).
- Since all modes are symmetric, allocation to a centre was done using Euclidean distance rather than the more expensive Mahalonobis distances.
- Once a centre had been found from the clustering procedure, a quasi-Newton optimisation was used to find the modal point for the local mode.

There is N times larger output from QuanTA than the PT version. So for comparison only a single randomly selected scheme from the QuanTA's output is used. Obviously, this should also be accounted for in any computational expense comparison.

With suitable tuning for the within temperature moves at the three stated temperature levels, both algorithms were run 10 times on this setup.

Figure 2.4 shows two representative trace plots of the target state chain for a run of the PT algorithm and QuanTA respectively. There is a clear improvement in the inter-modal mixing for the QuanTA scheme.

To further highlight this improvement, the consecutive swap acceptance rates between the three levels are given in Table 2.1. Clearly the rate of transfer of mixing information from the hot states to the cold state is significantly improved by the QuanTA scheme in this example.

Figure 2.5 further illustrates the inferential improvement "per iteration" of the QuanTA scheme over the standard PT scheme. Figure 2.5 compares the running modal weight approximation for the mode centred on 200 when using the standard PT and QuanTA schemes respectively. This used the cold state chains from 10 individual runs of the PT algorithm and 10 single schemes selected randomly from

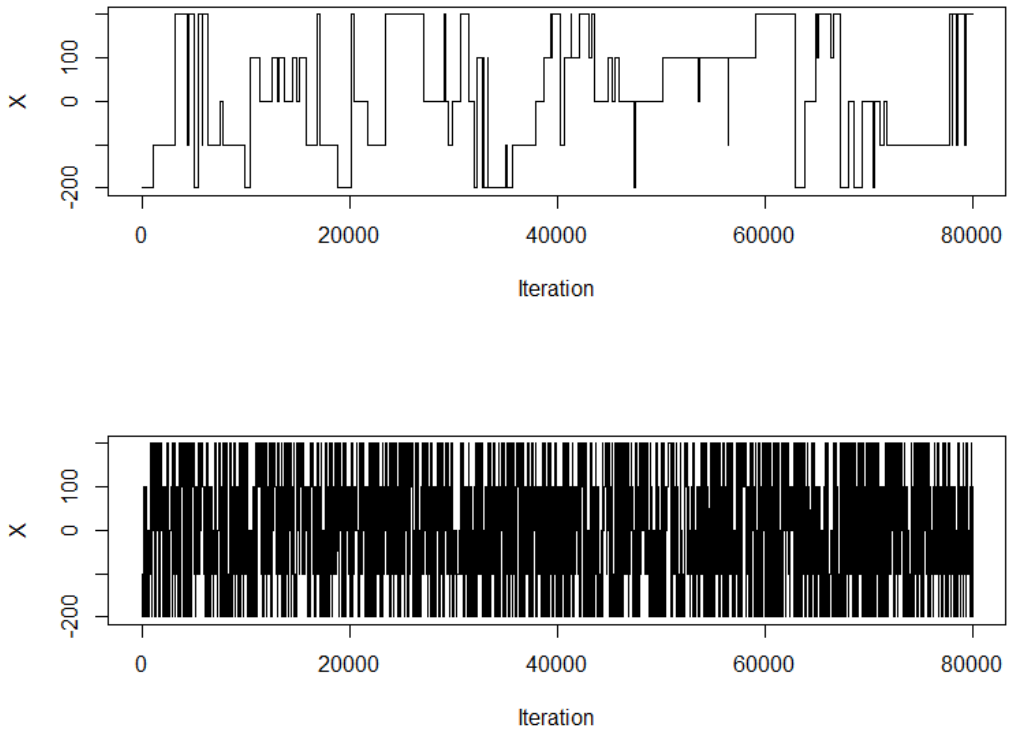

Figure 2.4: Trace plots of the target state chains for representative runs of the PT (top) and QuanTA schemes (bottom). Note the vastly improved inter-modal mixing of the new QuanTA scheme.

10 separate runs of the QuanTA algorithm. Then after removing a burn-in period of 2000 iterations of the chains, the running weight approximation of the mode centred on 200 was computed. Denoting the estimator of the k^{th} mode's weight by $\hat{w_k}$ and the respective cold state chain's i^{th} value as X_i and discarding a burn-in of period of B iterations,

$$\hat{w_k} = \frac{1}{N - B + 1} \sum_{i=B}^{N} \mathbb{1}_{\{c_k < X_i \leq C_k\}}. \tag{2.22}$$

where c_k and C_k are the chosen upper and lower boundary points for allocation to the k^{th} mode.

Figure 2.5 shows clearly that in this example the QuanTA scheme has a vastly improved rate of convergence to the true value of 0.2 over the PT scheme. The variability of the estimate after any finite number of iterations is visibly smaller for

Swap location:	1	2
PT	0.06	0.07
QuanTA	0.99	0.99

Table 2.1: Comparison of the acceptance rates of swap moves for the PT algorithm and QuanTA targeting the one dimensional distribution given in equation (2.21) and setup with the ambitious inverse temperature schedule given by $\{1, 0.0002, 0.0002^2\}$.

the QuanTA scheme. In fact, the bias from starting all schemes in the -200 centered mode is still not "forgotten" in the PT runs and has resulted in the majority of runs over this finite number of iterations significantly underestimating the weight w_5. In contrast the removal of a 2000 iteration burn-in appears to be ample for the QuanTA scheme, which shows fast, unbiased convergence to the true value of w_5.

For a computational expense comparison one should compare how many extra temperature levels would be required to make the PT scheme work optimally (i.e. with consecutive 0.234 swap acceptance rates). This gives a clearer idea of the reduction in number of intermediate levels that can be achieved using the QuanTA scheme and hence the savings in computational expense from running these extra chains. Using the same hottest state level of 0.0002^2 which ensures good hot state mixing over the statespace then repeated runs of the PT scheme were performed on temperature schedules that were geometrically generated (theoretically optimal in this setting) until a swap rate of approximately 0.234 was achieved between consecutive levels.

In this example a 0.04 geometric ratio suggested optimality for the PT scheme. Hence, to reach the stated hottest level needs 7 temperatures, as opposed to the 3 needed in the QuanTA scheme. Indeed, the common ratio for the spacings used in the QuanTA run was 200 times smaller.

2.7.2 Twenty-Dimensional Example

Again in the canonical setting of the Gaussian symmetric mixture distribution but now in a hard higher 20 dimensional case. The target distribution is a tri-modal Gaussian:

$$\pi(x) \propto \sum_{k=1}^{3} w_k \left[\prod_{j=1}^{20} \phi_{(\mu_k, \sigma^2)}(x_j) \right] \tag{2.23}$$

where $\phi_{(\mu,\sigma^2)}(.)$ is the density function of a univariate Gaussian with mean μ and variance σ^2. In this example, $\sigma = 0.01$, the marginal mode centres are given by $(\mu_1, \mu_2, \mu_3) = (-20, 0, 20)$ and all modes are equally weighted with $w_1 = w_2 = w_3$.

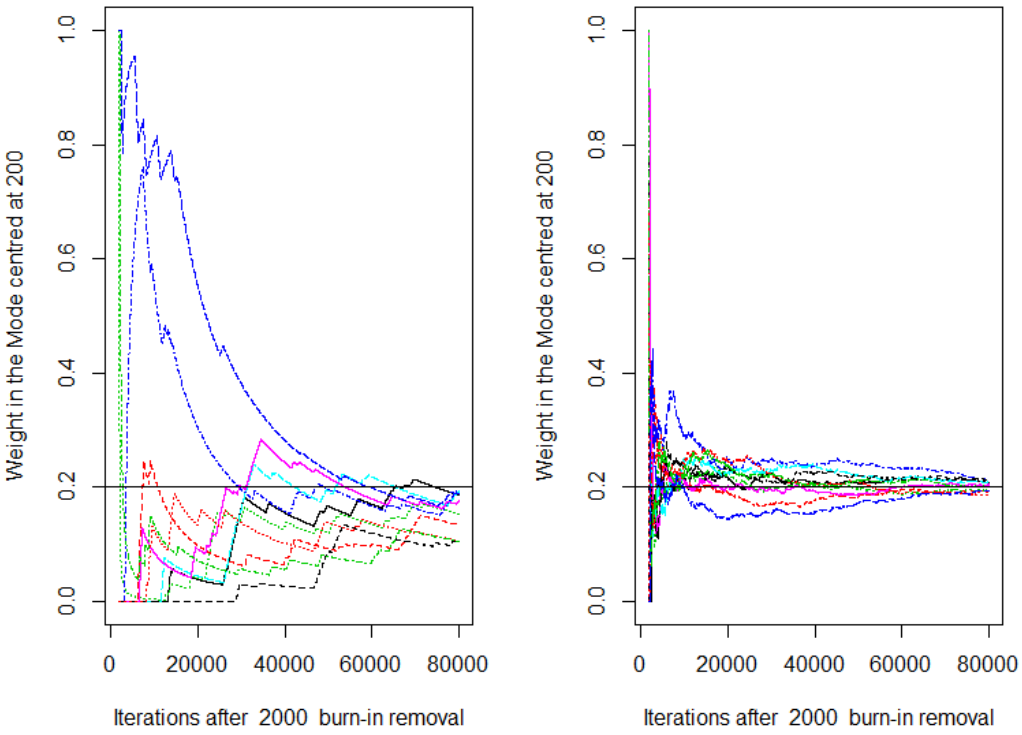

Figure 2.5: For the one dimensional target given in equation (2.21), the running weight approximations for the mode centred on 200 with target weight $w_5 = 0.2$ for 10 separate runs of the PT and QuanTA schemes respectively. Left: the PT runs showing slow and variable estimates for w_5. Right: the new QuanTA scheme showing fast, unbiased convergence to the true value for w_5

With such narrow well spaced modes in such a high dimension this is an extremely hard example for the PT algorithm and it will be seen that inter-modal mixing is very slow in contrast with the highly successful performance of the QuanTA in this case.

The temperature schedule for this example is derived from a geometric schedule (see Section 1.5.2) with an ambitious 0.002 common ratio for the spacings. Only 4 levels are used and so the temperature schedule is given by $\{1, 0.002, 0.002^2, 0.002^3\}$.

Both the PT and QuanTA schemes were run so that 20,000 swap moves would be attempted. For the QuanTA scheme this would be 20,000 swaps for each of the N schemes in parallel of which there were 100 in this example. Hence for a **single** scheme in each setting there is the same frequency and quantity of swap

proposals. Furthermore, in all runs **all** the chains were started from a start location of $(-20, \ldots, -20)$. This is biased towards the PT algorithm since it makes it hard for the clustering to find good initial clusters early on.

See the one-dimensional example for the specifics of the setup of the QuanTA scheme with regards the performance tuning parameters. With suitable tuning for the within temperature moves at the four stated temperature levels, both algorithms were run 10 times on this setup.

Figure 2.6 shows two representative trace plots of the target state chain for a run of the PT algorithm and QuanTA respectively. There is a clear improvement in the inter-modal mixing for the new QuanTA scheme. There is a stark contrast between the two algorithmic performances. The run using the standard PT scheme entirely fails to improve the mixing of the cold chain. In contrast the QuanTA scheme establishes a chain that is very effective at escaping the initialising mode and then mixes rapidly throughout the state space between the three separate modes.

The consecutive swap acceptance rates between the four levels are given in Table 2.2. Clearly there is no transfer of mixing information from the hot states to the cold state for the PT algorithm but there is effectively immediate transfer in the QuanTA scheme for this particular canonical example.

Swap location:	1	2	3
PT	0	0	0
QuanTA	0.99	0.99	0.99

Table 2.2: Comparison of the acceptance rates of swap moves for the PT algorithm and QuanTA targeting the Twenty dimensional distribution given in equation (2.23) and setup with the ambitious inverse temperature schedule given by $\{1, 0.002, 0.002^2, 0.002^3\}$.

Figure 2.7 compares the running modal weight approximation for w_3, the mode centred on $(20, \ldots, 20)$, when using the standard PT and QuanTA schemes respectively. This used the cold state chains from 10 individual runs of the PT algorithm and 10 single schemes selected randomly from 10 separate runs of the QuanTA algorithm. Removing a burn-in of 2000 iterations, the running approximation of w_3, i.e. $\hat{w_3}$, was computed. The weight approximation in each run was given by the estimator, described in the one dimensional example above, in equation (2.22). Due to the narrow and careful positioning of the modes on the hyper-diagonal of the $\mathbb{R}^{20}$ space then only the cold state, first component, one-dimensional chain was considered for this estimator. Furthermore, the chosen boundaries identifying the target mode were $c_k = 10$ and $C_k = \infty$.

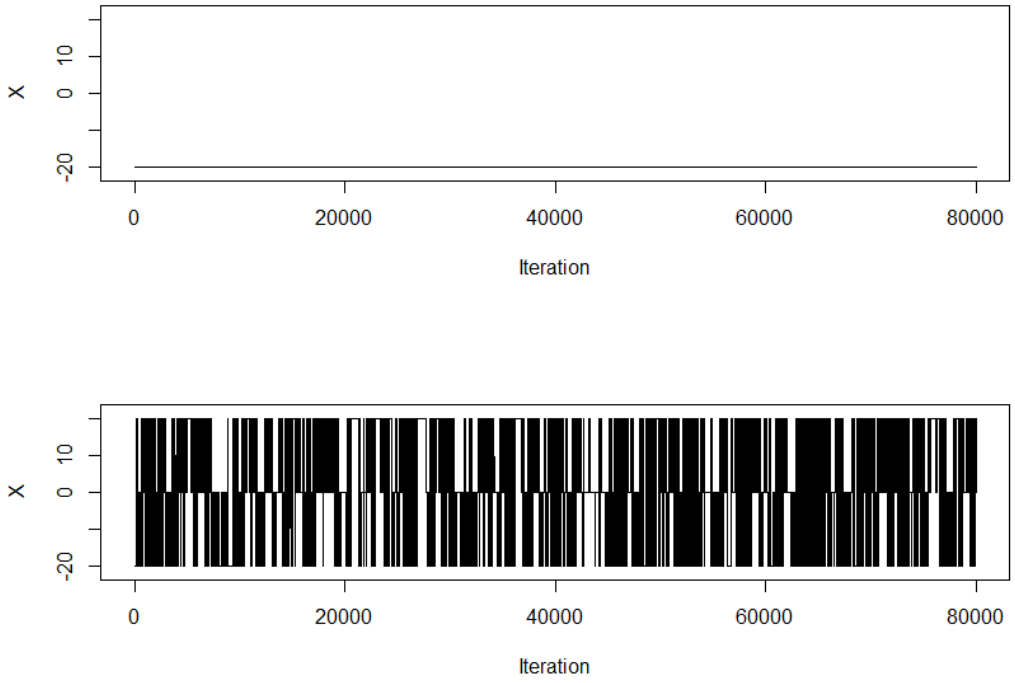

Figure 2.6: Trace plots of the first component of the twenty dimensional cold state chains for representative runs of the PT (top) and new QuanTA (bottom) schemes. Note the fast inter-modal mixing of the new QuanTA scheme, allowing rapid exploration of the target distribution. In contrast the (very boring) top plot shows that the PT scheme never manages to escape the initial mode marginally centred on -20 for the entirety of the run.

Figure 2.7 shows (predictably given the trace plots) that in this example the standard PT scheme entirely misses the mode in question and all 10 runs gave $\hat{w_3} = 0$; far from the true value $w_3 = 1/3$. In stark contrast, in the 10 runs of the QuanTA scheme, the 10 running estimators of w_3 quickly stabilise, with low variability about the true value.

A stark computational comparison is given by observing how many extra temperature levels would be required to make the PT scheme work optimally (i.e. with consecutive 0.234 swap acceptance rates). This gives an indication of the potential computational savings from running these extra chains. Using the same hottest state level of 0.0002^2 which ensures good hot state mixing over the state space then repeated runs of the PT scheme were performed on temperature schedules that

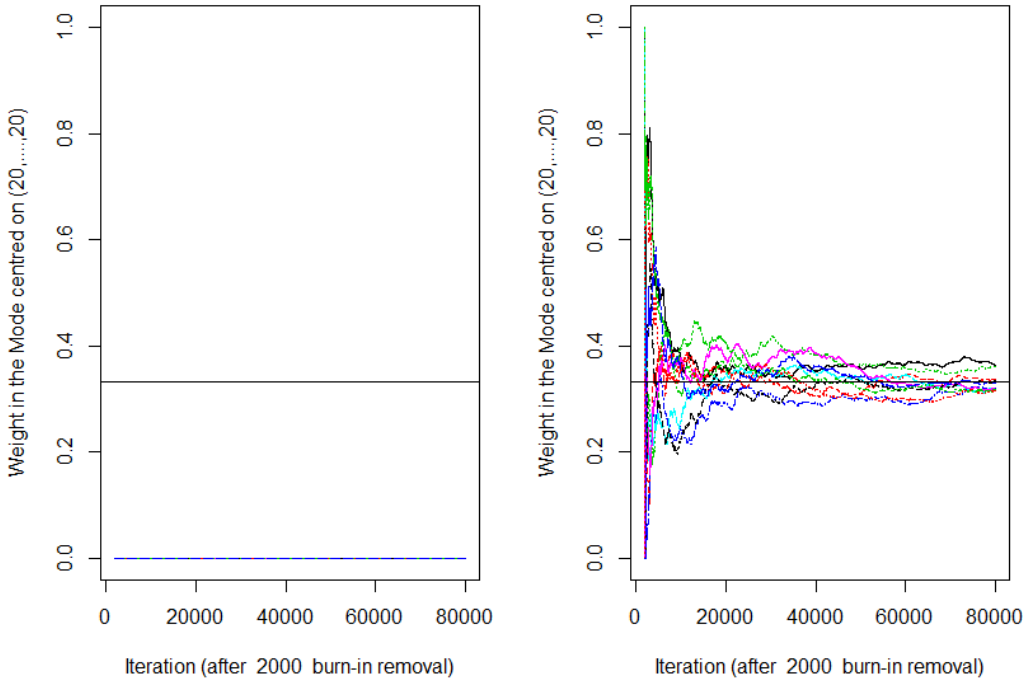

Figure 2.7: For the twenty-dimensional target given in equation (2.23), the running approximations for w_3 with target value 1/3 for 10 separate runs of the PT and QuanTA schemes respectively. Left: the PT runs confirmed that none of the 10 runs discovered the mode of interest. Right: the QuanTA scheme shows fast, unbiased convergence to the true value for w_3

were geometrically generated (theoretically optimal in this setting) until a swap rate of approximately 0.234 was achieved between consecutive levels.

In this example, simulations showed that a 0.58 geometric ratio induced swap acceptance rates of 0.234. Hence, optimal implementation of the PT approach would have required 36 temperature levels in contrast to the 4 that were sufficient for QuanTA in this example.

2.7.3 Five-Dimensional Non-Canonical Example

Leaving the canonical symmetric setting, the following example has a five dimensional Gaussian mixture target with even weight to the modes but with different

covariance scaling within each mode. The target distribution is given by:

$$\pi(x) \propto \sum_{k=1}^{3} w_k \left[\prod_{j=1}^{5} \phi_{(\mu_k, \sigma_k^2)}(x_j) \right] \tag{2.24}$$

where $\phi_{(\mu,\sigma^2)}(.)$ is the density function of a univariate Gaussian with mean μ and variance σ^2. In this example, $(\sigma_1, \sigma_2, \sigma_3) = (0.02, 0.01, 0.015)$, the marginal mode centres are given by $(\mu_1, \mu_2, \mu_3) = (-20, 0, 20)$ and all modes are equally weighted with $w_1 = w_2 = w_3$.

Although at first glimpse this doesn't sound like a significantly harder problem, or even far from the canonical setting, the differing modal scalings make this a much more complex example. This is due to the lack of preservation of modal weight through power-based tempering; a problem discussed and worked on in detail in Chapter 4. The K-means clustering procedure can be unstable and finds it hard to establish good mode centres early on in the run of the algorithm since the hot state chains are suggestive of the "wrong regions".

This was certainly the case in this example and it was evident that the clustering struggled badly when the differential modal scalings were made significantly harder than in this example. This shows that there is a lack of robustness with the transition phase of the algorithm before mode centres can become established. This reinforces that standard swap moves and perhaps methods from Chapter 4 should be used in tandem with this QuanTA scheme to add to the algorithm's robustness. Discussion of robustifying techniques can be found in Section 2.9.

The temperature schedule for this example cannot be a simple geometric schedule as in the previous example due to the scaling indifference between the modes. By using an ambitious geometric schedule, the clustering was very unstable early on and this often led to an inability to establish mode centres for the run. Instead, a mixture of geometric schedules was used with an ambitious spacing for the coldest levels and then a less ambitious spacing for the hotter levels. For the four coldest states an ambitious geometric schedule with 0.08 common ratio was used. A further 8 hotter levels were added using a conservative geometric schedule with ratio 0.4. Hence the schedule was given by:

$$\{1, 0.08, 0.08^2, 0.08^3, 0.4^9, 0.4^{10}, 0.4^{11}, 0.4^{12}, 0.4^{13}, 0.4^{14}, 0.4^{15}, 0.4^{16}\}. \tag{2.25}$$

For the QuanTA scheme, the reparametrisation moves were used for swap moves between the coldest 7 levels and standard swap moves were used otherwise.

Again, both the PT and QuanTA schemes were run so that 20,000 swap

moves would be attempted. For the QuanTA scheme this would be 20,000 swaps for each of the N schemes in parallel of which there were 100 in this example. Hence for a **single** scheme in each setting there is the same frequency and quantity of swap proposals. Furthermore, in all runs **all** the chains were started from a start location of 0.

Figure 2.8 shows two representative trace plots of the target state chain for a run of the PT and QuanTA algorithms respectively. There is a clear improvement in the inter-modal mixing for the QuanTA scheme; albeit far less stark than that in the canonical one-dimensional and twenty-dimensional examples already shown. There is still a stark contrast between the two algorithmic performances. The run using the standard PT scheme fails to explore the state space. The QuanTA scheme establishes a chain that is able to explore the state space but does appear to have a bit of trouble during burn-in; mixing is good therein.

The lack of comparable performance to the earlier canonical examples is entirely due to the lack of modal weight preservation when power tempering; an issue looked at in detail in Section 4. The runs still show impressive improvements in mixing for an ambitious spacing in the coldest part of the temperature schedule.

The consecutive swap acceptance rates between the 12 levels are given in Table 2.2. Clearly there is little transfer of mixing information through the 4 coldest states for the PT algorithm. Importantly, in the QuanTA scheme there are non-degenerate swap acceptance rates through the coldest levels but, unlike the canonical examples, they are not all close to 1. Indeed, for the QuanTA scheme, the swap rate between the coldest and second coldest levels is a relatively low (0.446) in comparison to the other swap rates for QuanTA. This is due to the "red-herring" effect caused by the lack of weight preservation in power-based tempering; and will be discussed in detail in Chapter 4.

This example is both positive (showing the improved mixing using the QuanTA scheme on a hard example) but also serves as a warning for the degeneracy of both the PT and new QuanTA schemes when using power-based tempering on a target outside of the canonical symmetric mode setting. Chapter 4 explores these difficult settings and begins to establish prototype methodology to deal with targets of the form given in this example.

2.7.4 Discussion of the Examples

It should be noted that the simulation examples are in the contrived settings of Gaussian mixture targets where the local Gaussianity of the modes is ideal for the operation of the algorithm. This is certainly an important class of distributions,

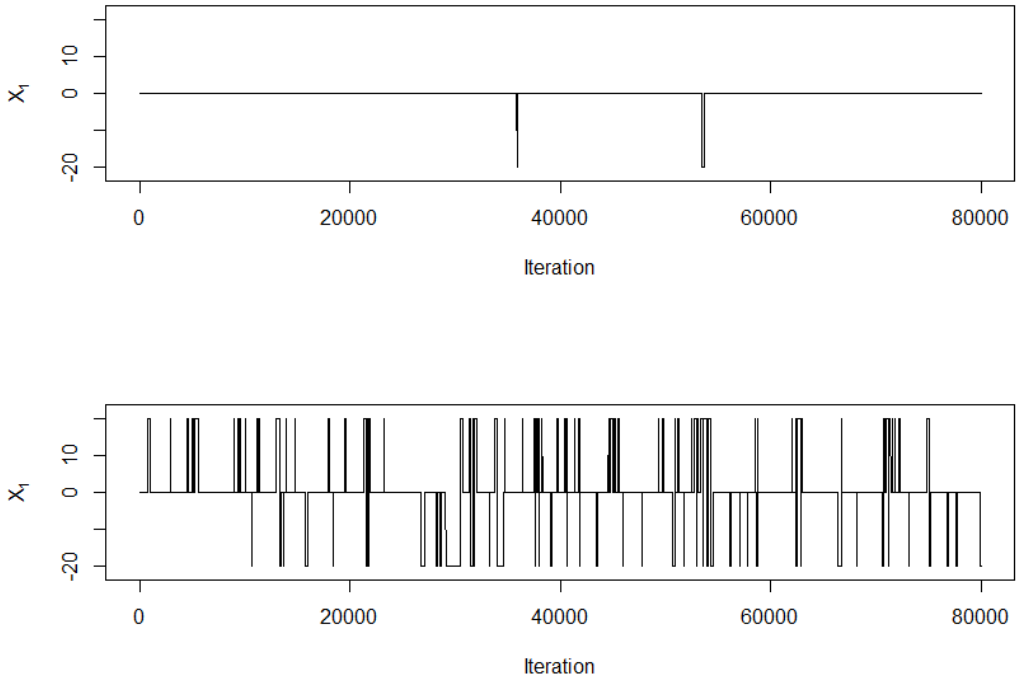

Figure 2.8: Trace plots of the first component of the five dimensional cold state chains for representative runs of the PT and QuanTA schemes respectively. Note the difference in inter-modal mixing between the QuanTA scheme and the PT scheme which (hardly) manages to escape the initial mode marginally centred on 0 for the entirety of the run.

with large data models giving rise to Bayesian central limit theorems to the local modes. Furthermore, the local Gaussian approximation required by QuanTA can always be made to any distribution which is second order continuously differentiable at the modal point. Indeed, the basis for a Gaussian type move is given further justification in Chapter 3. In particular in Theorem 3.4.1, which shows that with sufficient smoothness of the target distribution, QuanTA propels the mixing information through the temperature schedule at a "higher speed" than the PT approach.

Swap location:	1	2	3	4	5	6
PT	0.001	0.0161	0.0138	0.469	0.317	0.348
QuanTA	0.446	0.970	0.997	0.999	0.999	0.999
Swap location:	7	8	9	10	11	-
PT	0.328	0.334	0.359	0.324	0.327	-
QuanTA	0.285	0.285	0.285	0.285	0.302	-

Table 2.3: Comparison of the acceptance rates of swap moves for the PT and new QuanTA algorithm targeting the five dimensional distribution given in equation (2.24) and setup with the ambitious inverse temperature schedule given in equation (2.25). Note that for QuanTA, the reparametrised swap move was only used for swaps in the coldest 7 levels.

2.7.5 The Computational Cost of QuanTA

It is important to analyse the computational cost of the new approach. A basic analysis of the increase in computational expense of the new QuanTA approach relative to the traditional PT approach is undertaken. To be an effective algorithm the inferential gains of QuanTA per iteration should not be outweighed by the increase in run-time.

The analysis uses the runs of the one and twenty-dimensional examples, given above, using both the QuanTA and PT approaches. The algorithms were setup the same as in the ambitious versions of the spacing schedules in each case.

The key idea is to first establish the total run-time, denoted R, in each case. Typically one looks to compare the time-standardised Effective Sample Size (ESS). In this case it is natural to take the acceptance rate as a direct proxy for the effective sample size. This is due to the fact that the target distributions have symmetric modes with equal weights. Hence the acceptance rate between consecutive temperature levels dictates the performance of the algorithm; in particular the quality of inter-modal mixing.

To this end, taking the first level temperature swap acceptance rate, denoted A, the runs are compared using run-time standardised acceptance rates i.e. A/R.

Note that in both dimensional cases, the output from QuanTA is 100 times larger due to the use of 100 schemes running in parallel. Hence, for a standardised comparison the time was divided by 100. Therefore, in what follows in this section, when the run-time, R, of the QuanTA approach is referred to, this means the full run-time divided by 100. The fairness of this is discussed below.

In the one-dimensional example:

Algorithm	PT	QuanTA
Run-time (sec)	5.60	8.01
Swap Rate	0.06	0.99
A/R	0.01	0.12

In the twenty-dimensional example:

Algorithm	PT	QuanTA
Run-time (sec)	8.00	12.79
Swap Rate	0.00	0.99
A/R	0.00	0.08

In both cases the QuanTA approach has a longer run-time to generate the same amount of output; as would be expected due to the added cost of clustering. Indeed, it takes approximately 1.5 times longer to generate the "same amount of output".

However, the temperature swap move acceptance rates are 16.5 and ∞ times better respectively when using the QuanTA approach. Using the acceptance rate as a proxy for effective sample size then the quantity A/R is the most important value to compare. In both cases the QuanTA approach shows a significant improvement over the PT approach.

There are issues with the fairness of this comparison:

- By standardising the run-time of QuanTA by the number of parallel schemes is not entirely fair since it is sharing out the clustering expense between schemes. This is not fair since it implicitly assumes that a single run could have equally useful cluster centre approximations. Thus, this feature of the comparison favours QuanTA.

- The spacings are too ambitious for the PT approach meaning that the acceptance rates are very low. For a complete analysis one should run the PT algorithm on its optimal temperature schedule and then use the time-standardised ESS from each of the optimised algorithms.

The empirical computational studies are favourable to the QuanTA approach. This is for a couple of examples that are canonical for QuanTA. Outside of this canonical setting the improvements from running QuanTA will be less obvious. In

fact it will depend on the strength of the Gaussian approximation to the local mode, something looked at for the cold temperatures in Section 3.4.

Outside of the canonical symmetric mode setting then the acceptance rates will not necessarily be a suitable metric to compare mixing quality; indeed one would then need to consider an approach utilising ESS. However, this new algorithm is a prototype and so these computational results are encouraging.

2.8 Implications of Using the Reparametrisation Move in an Asymmetric Mode

This section assumes that after a mode centre has been established through the chosen clustering method then local optimisation is performed to find the mode point of the local mode. Consequently, reparametrisations would be centred about the mode point.

In a Gaussian mode, it is clear that when using the QuanTA reparametrisation move, the acceptance probability of the temperature swap move becomes independent of the position of the chain in the state space. This result is particularly reliant on the spherical symmetry of the mode.

It is important to understand the limitations of the algorithm and so it should be understood that the reparametrisation move will not be as powerful in the setting where the modes are not themselves symmetric.

To gain insight an asymmetric uni-modal "Gaussian" target will be considered. Consider a target given by the following (where $\phi_{(\mu,\sigma^2)}(.)$ is the density function of a Gaussian with mean μ and variance σ^2):

$$\pi(x) = \alpha \times \phi_{(0,\sigma^2)}(x)\mathbb{1}_{\{x<0\}} + \gamma \times \phi_{(0,1)}(x)\mathbb{1}_{\{x>0\}} \tag{2.26}$$

where α and γ are choosen so that $\frac{\alpha+\gamma}{2} = 1$ and $\alpha \times \phi_{(0,\sigma^2)}(0) = \gamma \times \phi_{(0,1)}(0)$ i.e. the mode is continuous in the first derivative. An example is given in Figure 2.9.

Denote the normalised tempered density at inverse temperature level β by π_N^β, and the normalisation constant of the tempered distribution at level β by $C(\beta)$. The acceptance probability of a temperature swap type move from level β to β' for a simulated tempering algorithm targeting π where the reparametrisation move towards the mode is used and without loss of generality the chain location, x, is in

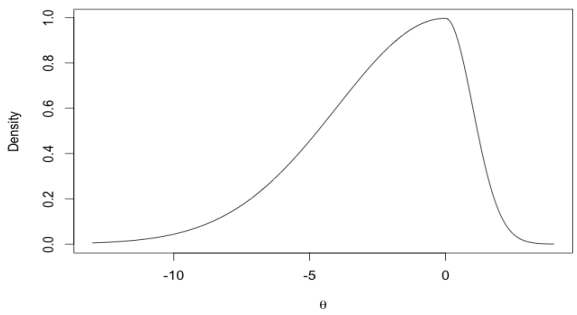

Figure 2.9: Example of an asymmetric Gaussian mode with density given by equation (2.26) and where $\sigma^2 = 16$, $\alpha = 10$ and $\gamma = 0.5$.

the upper tail of the distribution, is given by

$$\begin{aligned} \alpha &= \min\left(1, \frac{\pi_N^{\beta'}(x')}{\pi_N^{\beta}(x)} \left|\frac{\beta}{\beta'}\right|^{\frac{1}{2}}\right) \\ &= \min\left(1, \frac{C(\beta)}{C(\beta')} \frac{\gamma^{\beta'} \phi_{(0,1)}^{\beta'}\left(\left(\frac{\beta}{\beta'}\right)^{\frac{1}{2}} x\right)}{\gamma^{\beta} \phi_{(0,1)}^{\beta}(x)} \left|\frac{\beta}{\beta'}\right|^{\frac{1}{2}}\right). \end{aligned} \tag{2.27}$$

From the canonical true Gaussian setting then equation (2.27) can be simplified using

$$\frac{D(\beta)}{D(\beta')} \frac{\phi_{(0,1)}^{\beta'}\left(\left(\frac{\beta}{\beta'}\right)^{\frac{1}{2}} x\right)}{\phi_{(0,1)}^{\beta}(x)} \left|\frac{\beta}{\beta'}\right|^{\frac{1}{2}} = 1 \tag{2.28}$$

where $D(\beta) = \int_{-\infty}^{\infty} \phi_{(0,1)}^{\beta}(z)dz$.

So by using the result in equation (2.28) then

$$\begin{aligned} \alpha &= \min\left(1, \frac{\gamma^{\beta'} D(\beta')/C(\beta')}{\gamma^{\beta} D(\beta)/C(\beta)}\right) \\ &= \min\left(1, \frac{\gamma^{\beta'}\left[\int_0^{\infty} \phi_{(0,1)}^{\beta'}(z)dz\right]/C(\beta')}{\gamma^{\beta}\left[\int_0^{\infty} \phi_{(0,1)}^{\beta}(z)dz\right]/C(\beta)}\right) \\ &= \min\left(1, \frac{\text{Proportion of mass in the upper tail at level } \beta'}{\text{Proportion of mass in the upper tail at level } \beta}\right). \end{aligned} \tag{2.29}$$

In the asymmetric case the acceptance probability is no longer 1. In modes that are of this form (or at least approximately) then equation (2.29) shows that, conditional on the tail side of the mode, the acceptance probability of a swap move becomes independent of the location of the chain. So there is still the nice interpretation that, conditional on the tail, the swap acceptance probability is independent on the location. Obviously this is only approximately true in the case that the tails are approximately Gaussian.

However, this illustration is in a single dimension. Supposing a target consists of a product of iid such components, then as the dimension grows then, conditional on all components being in the same tail marginally, this acceptance ratio will decay (or grow) geometrically fast in dimension. This would then require a finite spacing in the temperature schedule rather than the arbitrary spacing for the canonical Gaussian setting.

Interestingly, the aforementioned work analysing the utility of the QuanTA approach in high-dimensional super cold temperatures in Chapter 3 in Theorem 3.4.1 actually concludes that there is a reduced higher order behaviour for the temperature spacings when there is asymmetry about the mode point.

2.9 Auxiliary Cold Levels Aiding the Performance of the QuanTA Weighted Clustering

In all the examples considered the reparametrised shift has been centred about the mode point of the local mode. The proposal has been that the weighted K-means procedure locates a mode centre which will at least be contained in the "basin" of the local mode and then localised optimisation perhaps using gradient ascent, (stochastic gradient ascent if the data size of the problem is large e.g. Bottou [2010]) or a Newton optimisation. Only needing the centre point to be located in the basin of attraction of the local mode allows a reduction in the number of parallel schemes due to reduced need for accuracy of the mode point estimate. However, there may be scenarios where one may not want to employ optimisation schemes but still want a reasonable and stable estimate of the mode points.

Consider the process of optimisation via simulated annealing optimisation, Kirkpatrick *et al.* [1983], which is heavily linked to parallel tempering, and has the aim of discovering a global maximum. It does this by slowly cooling (as opposed to heating) the target distribution until the chain has been trapped in the region about the global maximum (i.e. the mode point of the local mode).

For the standard PT algorithm there is no obvious gain to go colder than

the cold target state (beyond perhaps storing history to provide memory of old locations similar to the motivation in Brooks *et al.* [2003]). Indeed having the colder temperatures would add to the computational complexity and run time of the algorithm if one hasn't parallelised and even then it is unclear that the added cost of mixing through these auxiliary super cold temperatures would be worth the benefits of history storage. This analysis is left for further work. However, in the setting of the QuanTA algorithm there is motivation to using these colder states.

For particles that are in the discovered/explored modes then if particles in these modes have reached the colder states then they are less disperse about the local mode than those in the hotter temperatures and thus have more precision in estimating the **mode point**. Hence, using these colder states could improve the stability of the cluster centre values. Indeed, if the target is C^2 then a Laplace approximation to the local mode becomes increasingly accurate as the temperature gets colder, see Section 3.4.1. This would mean that the chains at these colder levels should be able to mix through the additional colder levels increasingly easily.

Trial runs have shown that using colder levels in conjunction with the weighted K-means setup, given in Section 2.5.1, gives an algorithm that has improved stability to the local mode approximation. Additionally, these auxiliary super-cold levels are increasingly centred about the **mode point** rather than the **mean** of the local mode which could be useful in an asymmetric modal setting.

Empirical Example:

Consider the toy uni-variate bi-modal Gaussian target distribution

$$\pi(x) = \frac{1}{2}\phi_{(-\mu,\sigma^2)}(x) + \frac{1}{2}\phi_{(\mu,\sigma^2)}(x)$$

where $\mu = 50$ and $\sigma = 4$. Albeit in the canonical setting, the modes are highly dispersed even at the target cold temperature. This would mean that for stable estimates of the mode point even via the weighted cluster centre approach one would need a large number of parallel schemes to provide a large collection of particles.

So consider running the QuanTA algorithm with weighted clustering to target this distribution using the following temperature schedules

1. A standard schedule where the coldest state is the target level, $\beta = \{1, 0.25, 0.05, 0.005\}$.
2. A schedule including temperatures that are far colder than the target level, $\beta = \{20, 10, 1, 0.25, 0.05, 0.005\}$.

Clearly, if more particles are used in parallel then the mode centre approximations will be less variable. In fact, with local mean calculated in equation (2.19), derived from the MLE for a unimodal Gaussian, then one would expect an error of order $O(1/n^{\frac{1}{2}})$ assuming n particles are used. Hence, in an attempt to make a "fair" comparison of mode centre approximation stability between the different runs then the computational budget should remain constant and thus the same number of particles used. Each version was therefore run with a different numbers of parallel versions in such a way that the number of particles used in the clustering procedures were equal. In this comparison, 100 versions were run for the standard schedule and 60 versions were run for the super cold version. In both cases only particles at inverse temperature level 0.05 and colder were used for clustering, thus ensuring identical clustering complexity in terms of the number of particles used.

In the competing runs of the algorithm the mode centre approximations were recorded and a burn-in period removed. The distribution of the estimated cluster centres for the mode located at 50 was analysed. Figure 2.10 compares the kernel density estimates of the distributions of the estimated centres in the two cases. It is clear that the super cold temperatures improve the stability of the estimates, with the simulations giving standard deviations of 0.71 and 0.18 for the standard and super-cold schedule versions respectively.

It is evident that using these super-cold temperatures can add robustness to the algorithm. In cases where the modes are asymmetric then using the super cold temperatures will have densities concentrated about the mode point. Hence, using weighted clustering amongst these super cold levels will give better estimates of the mode points (prior to any local optimisation).

Some general comments on using the super cold auxiliary levels:

- If using super-cold temperatures, care should be taken to establish fully weighted clustering only once the population has reached invariance otherwise it may be hard to establish new modes away from those where the super cold states are initiated. One suggestion is to sporadically use un-weighted clustering in an attempt to establish new modes and add robustness.

- When the target is suitably smooth then QuanTA can mix very fast through the super-cold temperatures meaning that the spacings can be ambitious and not as many levels would be required, see Theorem 3.4.1.

- Once at invariance, providing there are enough particles in parallel, then clustering could be restricted to just those locations at the coldest temperatures

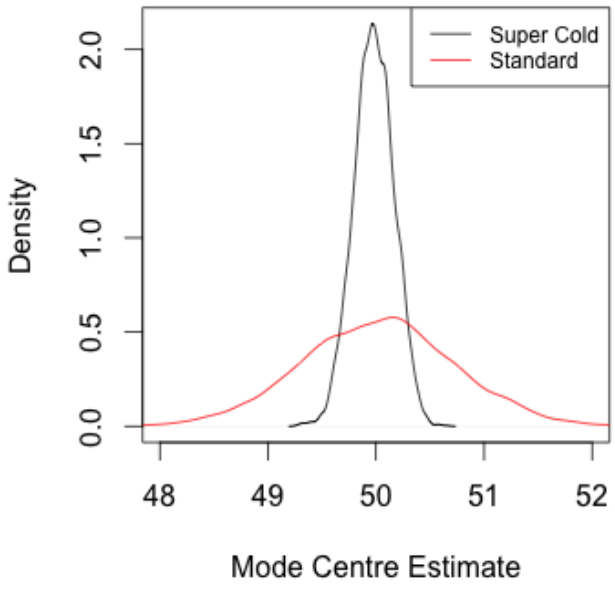

Figure 2.10: Plot of the kernel density estimates of the distributions of the mode centre estimates for the mode centred on 50 for both the version of the QuanTA algorithm using a standard temperature schedule and then using a temperature schedule with super cold levels. It is clear that the super cold version gives more stable estimates with much less variability.

reducing the computational overhead on clustering and providing accurate mode point approximations.

2.10 Robustification in Non-Gaussian cases

So far the focus has been on the canonical Gaussian type modes. This canonical setting covers an important class of models. However, Gaussianity certainly doesn't cover all the modal structures. The deterministic form of the reparametrisation move is specific and potentially restrictive. This section explores some basic ways to establish a more robust version of the algorithm that may be of interest to a practitioner and be more widely applicable.

2.10.1 Alternative Reparametrisations

Recall the following, regarding the preservation of quantiles in general for a simulated tempering swap-type move. Suppose that the current location of the chain in a simulated tempering algorithm is x at inverse temperature is β and that a temperature swap move $\beta \to \beta^{'}$ is proposed along with a reparametrised deterministic shift of $x \to x^{'} = g(x, \beta, \beta^{'})$. Now suppose that g is chosen in a way that preserves the quantile of the target, π, at the respective temperature levels. Denoting the CDF of π^{β} by F_{β} then preservation of the quantile requires g such that

$$F_{\beta}(x) = F_{\beta'}(g(x, \beta, \beta^{'})) \tag{2.30}$$

and so by differentiating wrt x and rearranging gives

$$1 = \frac{\pi^{\beta^{'}}(g(x, \beta, \beta^{'}))}{\pi^{\beta}(x)} \left| \frac{\partial g(x, \beta, \beta^{'})}{\partial x} \right|$$

which is exactly the acceptance ratio in the temperature swap move for the simulated tempering move. Unfortunately this doesn't give a general approach to making all swap moves have acceptance probability 1. For reversibility to hold one needs to ensure that the function g is a bijection and so has an inverse. Solutions to equation (2.30) are certainly not unique in a general multidimensional non-trivial distribution and so it is not a good idea to solve numerically as it is unlikely to give reversible results.

In the Gaussian case however, there is the reparametrisation move which is given by a linear (and therefore bijective) function of the location. As the shift in this canonical setting solves equation (2.30) and is also invertible (providing reversibility) then as a consequence acceptance rates of the swap move are maximised. Particularly important to note is that the shift is performed componentwise and thus is effectively preserving each of the marginal quantiles and so the correlations between components has no effect on the reparametrisation.

The following section explores toy scenarios when a similar reparametrisation based shift can be attempted and the Gaussian type shift would not work well.

2.10.2 Examples of Reparametrisations in Important Non-Gaussian Cases

The first thing to note is that the extension to non-Gaussian modes will focus on cases where the d-dimensional modes are of an iid component form (or at least have

spherical contours) and thus the preservation of the marginal quantiles is sufficient to preserve the full quantile of the joint distribution and importantly provide solutions that are reversible (since univariate marginal preservation gives a bijective reparametrisation).

Focus will be given to three examples that encapsulate the most common behaviours of cases when the modes are heavier tailed than the Gaussian. These are the Laplace, scaled t and a more general polynomially tailed mode which has a more tractable solution than the t distribution.

The Laplace Distribution:

In the case of the Laplace distribution where the pdf is given by

$$\pi(x) = \frac{1}{2}\lambda \exp\left(-\lambda|x|\right) \quad \text{for} \quad x \in (-\infty, \infty) \tag{2.31}$$

then in a simulated tempering context the ideal reparametrisation when proposing a move from inverse temperature level β to $\beta^{'}$ is to make the deterministic reparametrisation $x \to \frac{\beta}{\beta'}x$. Note the difference from the canonical Gaussian version where the ideal is to make the reparametrisation $x \to \left(\frac{\beta}{\beta'}\right)^{1/2} x$. It is clear that, in the Laplace case, the QuanTA shift will never coincide with the optimal Laplace shift even for super-cold temperatures (since the laplace distribution is not C^2 at it's mode). Using the Gaussian-based shift in this setting gives some improvement in the swap acceptance probability since it is moving the particles locations in the "right direction" marginally but inevitably for this heavier tailed target "not doing enough".

The Scaled t distribution:

A t-distribution with a large degrees of freedom will essentially be Gaussian, particularly about the mode point, and so the QuanTA scheme will work well in such examples.

For a t-distribution when the degrees of freedom are small and thus the distribution is particularly heavy tailed, then as for the Laplace distribution example, the Gaussian shift is simply not significant enough to preserve the modal quantiles between temperature levels.

The density function for a standard scaled $t_{(\nu,\sigma)}$ distribution is

$$\pi(x) \propto \left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \tag{2.32}$$

where ν is the degrees of freedom and σ the scale parameter. Indeed, even after tempering this distribution to inverse temperature β it remains a scaled t distribution such that $\pi^{\beta} \sim t_{(\nu',\sigma')}$ where $\nu' = \beta\nu + \beta - 1$ and $\sigma' = \sigma\nu^{1/2}\left(\beta\nu + \beta - 1\right)^{-1/2}$.

Suppose the target of a simulated tempering algorithm is the $t_{(4,1)}$ and that the chain is currently at temperature level $\beta = 0.5$ and a swap move of the form $\beta \to \beta' = \beta + \epsilon$ is proposed. Furthermore, suppose that the current position of the chain is at $x = 2$. Figure 2.11 compares x', for the Gaussian motivated shift versus the implicitly calculated ideal quantile location that would preserve the quantile as the value of β' varies.

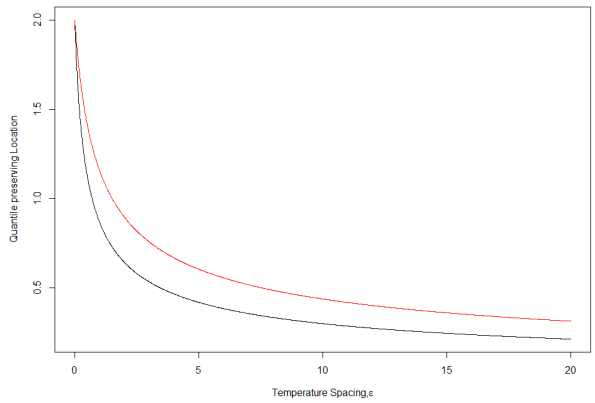

Figure 2.11: Locations of x', for the Gaussian motivated shift (red) versus the true quantile preserving location (black) when the target of a simulated tempering algorithm is the $t_{(4,1)}$ and that the chain is currently at temperature level $\beta = 0.5$ and a swap move of the form $\beta \to \beta' = \beta + \epsilon$ is proposed with the chain currently located at $x = 2$. This is plotted over the range of values for the spacing, $\epsilon \in [0, 20]$. It shows that although the Gaussian based shift is going in the "right direction" it never shifts far enough to preserve the quantile.

Even in the case where one assumes knowledge of the parameters σ and ν for the scaled t, the location of the reparametrised chain is not analytic and must instead be computed implicitly from the one-dimensional marginal CDFs. The following tractable heavy tailed example gives a starting point for establishing an analytic reparametrisation since it has identical tail behaviour to a t-distribution.

A Polynomially Tailed Distribution:

Consider a target distribution with mode point μ and density function given

by

$$\pi(x) = \frac{(k-1)}{2\sigma}\left(1 + \left|\frac{(x-\mu)}{\sigma}\right|\right)^{-k} \qquad \forall\ x \in \mathbb{R}. \tag{2.33}$$

Now suppose that the chain in a simulated tempering algorithm is at location x at inverse temperature level β and that a temperature swap proposal is made such that $\beta \to \beta^{'}$. Then by computing the CDF and subsequently the location that one should shift to preserve the quantile then one can show that

$$x^{'} = \begin{cases} \sigma\left[1 - \left(1 - \frac{(x-\mu)}{\sigma}\right)^{\frac{k\beta-1}{k\beta^{'}-1}}\right] & \text{if } x \leq \mu, \\ \\ \sigma\left[\left(1 + \frac{(x-\mu)}{\sigma}\right)^{\frac{k\beta-1}{k\beta^{'}-1}} - 1\right] & \text{if } x > \mu. \end{cases} \tag{2.34}$$

Note that in the limit as $\beta \to \infty$ then the required transformation converges to that of the Laplace distribution type reparametrisation.

There are certainly two major complications with using this reparametrisation that make it particularly hard to implement effectively:

1. The reparametrisation formula depends upon knowing the value of (the more than likely) unknown parameter k

2. The reparametrisation formula depends upon knowing the value of (the more than likely) unknown parameter σ.

Now this is significantly more information than was required for Gaussian-motivated QuanTA version. In the population-based setting of QuanTA then there is scope for estimation of the two parameters from the clustered sample in the mode. However, with sample sizes potentially low and the nature of the heavy tails, then maximum likelihood estimates could be very poor and unstable. An initial suggestion would be to assume knowledge of the parameter k and estimate σ by using the method of moments. This could be done by noting that for a particle, x, at inverse temperature level β then by making the transformation

$$y = \left(\frac{(k\beta-2)(k\beta-3)}{2}\right)^{\frac{1}{2}}(x-\mu) \tag{2.35}$$

then the variance of the new variable, y, is σ^2. The stability and practicality of this approach is beyond the scope of this project.

The question of how to "guess" k is then a further issue. However, there are some key considerations to take into account. If $k\beta \leq 3$ then the second moment

doesn't exist and so the above approach to estimate σ would be invalid. More seriously, if $k\beta \leq 1$ then the distribution is improper and hence the algorithm wouldn't be targeting a well defined invariant distribution at the hotter levels.

Ultimately, in a multi-modal setting with heavy-tailed distributions one would require a global reparametrisation to the problem to obtain exponential tails for the modes, otherwise beyond a certain hot temperature the tempered distributions would become improper since their integrals will be infinite.

An important question for the practitioner is whether the QuanTA algorithm can be made more robust to work in different modal cases. The QuanTAR algorithm introduced in this following Section 2.10.3 suggests a basic solution to add some level of robustness to QuanTA but this comes at a cost to the efficiency.

2.10.3 Robustification in the Non-Gaussian Modes: The QuanTAR Algorithm

An attempt to robustify the QuanTA algorithm that fits naturally into the QuanTA algorithm's framework is motivated by the random scan Gibbs sampler. If the different reparametrisation moves that the user wants to use are all considered as different proposal kernels then every time a temperature swap move in the algorithm is proposed then the form of the reparametrisation move is chosen at random from the set of all possible moves available. These choices could be given weights so that if the user wants predominantly the canonical Gaussian move then extra weight could be added to performing such moves.

By randomising the selection of reparametrisation move then there will be a loss of efficiency from doing the optimal moves every time. Denote by P_i the i^{th} reparametrisation type move for $i \in 1, \ldots, T$, and suppose that the target is a unimodal Gaussian. If all move types are chosen uniformly then it would be expected that the algorithm would be at worst $1/T$ times as "efficient" as the ideal version of the algorithm which would only use the Gaussian reparametrisation.

This potential loss of efficiency could actually be integral to adding robustness in cases where the temperature spacings are ambitious with inhomogeneous modal distributions. In this case (in the parallel tempering setting in the QuanTA algorithm) the robustified version would have a chance of doing the ideal move every time a swap move is proposed.

As such, and mostly for completeness, a toy suggestion for a more robust version of QuanTA is presented. This is called the QuanTAR (Quantile Tempering Algorithm-Robust)

The QuanTAR Algorithm:

Denote by P_i the i^{th} reparametrisation type move for $i \in 1, \ldots, T$ and the corresponding weights for each move by w_i respectively. Furthermore, denote by $g_P(x, \beta_1, \beta_2, \mu)$ the transformation, of type P, of location x that is the location of the chain in the state space at the tempering level β_2 and in the local mode centred on μ and is being swapped to level β_1. As in QuanTA, N parallel schemes are run in parallel. Then QuanTAR is as follows

- Choose a sequence of tempering values $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$.
- Choose initial values of the chains for each temperature level, $x^k_{00}, x^k_{01}, \ldots, x^k_{0n}$ for each k of the N parallel schemes.
- Choose the proposal mechanism for a given within temperature move, $q_{\beta_j}(x^k_{ij}, x^k_{(i+1)j})$ for $j = 1, \ldots, n$.
- Choose the number, m, of within temperature proposals the chains will perform before attempting a swap type move and choose the total number, s, of swap moves that will be attempted.
- After running the chains in parallel for a burn-in period, iterate s times:
  1. Perform a clustering procedure (e.g. k-means) on the locations of the particles for the particles x^k_{ij} where $k \in \{1, 2, \ldots, N/2\}$ and $j \in \{1, 2, \ldots, y\}$ and i the most recent iteration of the chain.
  2. For each $k \in \{N/2, \ldots, N\}$ a swap move for the k^{th} scheme is proposed. This is done as follows:
     - i Uniformly randomly select a pair of adjacent temperatures, $1/\beta_j$ and $1/\beta_{j+1}$ say, for which a reparametrised swap move will be proposed, and where the values of the respective chains are (currently) x^k_{ij} and $x^k_{i(j+1)}$. If the tempering level is too high i.e. $(j+1) > y$ then propose the standard swap move and accept with the ratio given in equation (2.36). Otherwise continue with the reparametrisation move proposal.
     - ii Classify the clusters to which the particles x^k_{ij} and $x^k_{i(j+1)}$ belong, denoted by μ_1 and μ_2 respectively.

iii Choose $m_1, m_2 \in \{P_1, \ldots, P_T\}$ randomly such that $\mathbb{P}(m_1 = P_i) = w_i$ and $\mathbb{P}(m_2 = P_j) = w_j$ independently.

iv The transformed locations are computed for the reparametrisation move using the respectively randomised move types x and y i.e.

$$g_{m_1}(x_{ij}^k, \beta_{(j+1)}, \beta_j, \mu_1) \text{ and } g_{m_2}(x_{i(j+1)}^k, \beta_j, \beta_{(j+1)}, \mu_2)$$

and also the respective Jacobians

$$J_1 = \frac{\partial g_{m_1}}{\partial x}(x_{ij}^k, \beta_{(j+1)}, \beta_j, \mu_1) \text{ and } J_2 = \frac{\partial g_{m_2}}{\partial x}(x_{i(j+1)}^k, \beta_j, \beta_{(j+1)}, \mu_2).$$

v Compute the acceptance ratio for the proposed swap and accept the swap with probability equal to

$$\min\left(1, \frac{\pi(g_{m_1}(x_{ij}^k, \beta_{(j+1)}, \beta_j, \mu_1))^{\beta_{(j+1)}} \pi(g_{m_2}(x_{i(j+1)}^k, \beta_j, \beta_{(j+1)}, \mu_2))^{\beta_j} |J_1||J_2|}{\pi(x_1)^{\beta_1} \pi(x_2)^{\beta_2}}\right).$$

3. Now perform the clustering procedure (e.g. k-means) on the locations of the particles for the particles x_{ij}^k where $k \in \{N/2, \ldots, N\}$ and $j \in \{1, 2, \ldots, y\}$ and i the most recent iteration of the chain.
4. Repeat the procedure in step 2 but now for $k \in \{1, 2, \ldots, N/2\}$.
5. For each of the N parallel schemes, perform m within temperature moves for each of the $(n+1)$ chains according to the proposal mechanism specified.

Toy Example of the QuanTAR Algorithm:

Consider the univariate bi-modal target distribution which is a mixture of a Laplace distribution and a Gaussian given by

$$\pi(x) = 0.5 \times \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}(x+100)^2\right) + 0.5 \times \frac{1}{2}\exp(-4|x-100|) \ \ \forall x \in \mathbb{R}. \quad (2.36)$$

Both the standard parallel tempering and QuanTAR algorithm were run to target this distribution on an ambitious geometric temperature schedule given by $\beta = \{1, 0.1, 0.01, 0.001\}$, with no use of auxiliary super cold temperatures. This leads to a series of targets illustrated in Figure 2.12.

The reparametrisation options for a swap mode were chosen uniformly from the set of options only containing the Gaussian and Laplace motivated moves (clearly

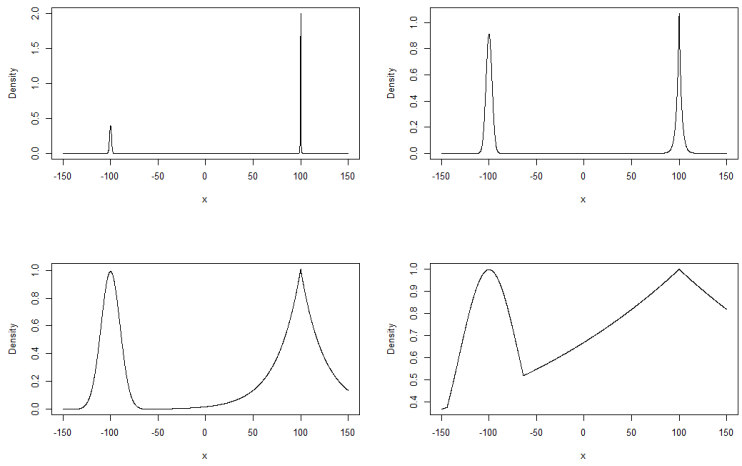

Figure 2.12: Plots showing the target density, π, given in equation (2.36), at each of the four temperature levels used in the implementation of the QuanTAR and PT algorithms, $\beta = \{1, 0.1, 0.01, 0.001\}$.

making this a very contrived example). The reparametrisation move was only used for swaps between the coldest two temperature levels.

The acceptance rate for the swap move between the coldest two temperature levels is improved when using the QuanTAR algorithm over the standard PT algorithm, 0.57 and 0.27 respectively. Figure 2.13 shows the trace plots of the chains located in the coldest temperature in both the QuanTAR and PT runs. The inter-modal jump rate is higher in the case of the QuanTAR run, and hence the increased acceptance rates are indeed corresponding to an increased rate of mixing between modes.

The increased rate of mixing in this example is much harder to notice from the trace plots than in the examples for the QuanTA approach. Much of this is due to the loss of efficiency from using the random scan selection of the reparametrisation moves. Also significant is a major issue regarding the modal weight preservations. From the density plots in Figure 2.13 there is a lack of modal weight preservation when using power-based tempering, as was the case in the five-dimensional QuanTA example earlier. This is a major flaw of power-based tempering and this is the focus of the work in the following two chapters.

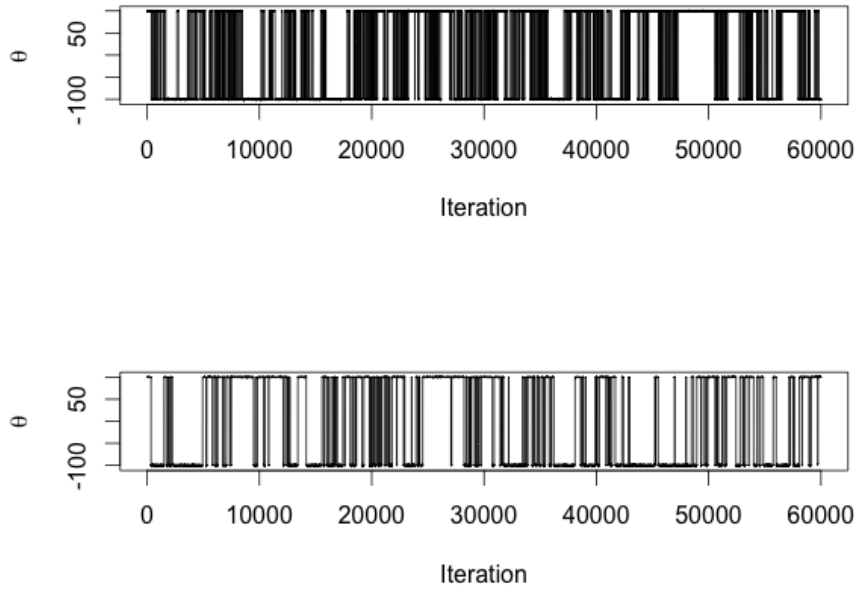

Figure 2.13: Denoting the sample values as θ instead of x. Top: Trace plot of the cold state chain in the QuanTAR run. Bottom: Trace plot of the cold state chain in the PT run.

Chapter 3

Optimal Scaling of the QuanTA Algorithm

3.1 Introduction

Chapter 2 established that in the canonical Gaussian unimodal setting, QuanTA can make arbitrarily large steps through the temperature schedule. However, for a general setting, the move will have limitations to the spacing ambitiousness as there was for the PT algorithm. This section will establish a similar result to Atchadé *et al.* [2011], giving practitioners a gauge on the optimal setup for QuanTA with relation to the consecutive temperature spacings. It will be shown that a consecutive spacing that induces a swap rate of approximately 0.234 is optimal according to the metric used in Atchadé *et al.* [2011].

Atchadé *et al.* [2011] motivated seeking an optimal temperature schedule selection for the efficiency of the transfer of the hot state mixing information through to the cold state for the standard PT algorithm. One measure of the transfer efficiency through the temperature schedule is the Expected Squared Jumping Distance, $ESJD_\beta$, for a temperature swap move. This is used as the metric in Atchadé *et al.* [2011] and is given added justification as the quantity which limits to the total variation for the limiting diffusion (if one indeed exists) of the process, Roberts and Rosenthal [2014]. Denote by $ESJD_\beta$ the expected squared jumping distance for the expectation of the square of the difference in inverse temperature change for a chain at inverse temperature level β undertaking a proposed swap with a chain at a colder temperature level $\beta' = \beta + \epsilon$. Mathematically that is

$$ESJD_\beta \quad = \quad \mathbb{E}\left[(\gamma - \beta)^2\right] \tag{3.1}$$

where β is the current temperature of the chain and γ is the random variable taking the the value β if the proposed swap move is rejected or $\beta^{'}$ if the move is accepted.

In order to pass the information efficiently from the hot state to the cold state then one needs a strategy to balance making overly ambitious large jump proposals which have low acceptance probabilities against under ambitious small jump proposals with overly high acceptance rates; both of which lead to slow mixing. By tuning the consecutive temperature spacings to maximise the $ESJD_\beta$ between levels then a strategy balancing ambition and acceptance should be reached.

3.2 The Setup and Theorem Statement

Consider a d-dimensional target distribution of the very simple form

$$f_d(\mathbf{x}) = \prod_{i=1}^{d} f(x_i) \tag{3.2}$$

where the marginal distributions f is in C^4 with a global maximum at the point μ. Furthermore, the marginal targets f are assumed to be of the form

$$f(x) = e^{-H(x)} \quad \forall x \in \mathbb{R} \tag{3.3}$$

where the $H(x) := -\log(f(x))$ is regularly varying i.e. there exists an $\alpha > 0$ such that for $x > 0$

$$\frac{H(tx)}{H(t)} \to x^\alpha \text{ as } |t| \to \infty. \tag{3.4}$$

This is a sufficient condition for Theorem 3.2.1 and ensures the moments and integrals required for the proof are all well defined. In addition to this assume that the fourth derivatives of $(\log f)(\cdot)$ are bounded, i.e. $\exists M > 0$ such that

$$|(\log f)''''(z)| < M \quad \forall z \in \mathbb{R}. \tag{3.5}$$

This condition is sufficient for proving Theorem 3.2.1 but is far from necessary. In fact, the proof should still work if the condition is weakened so that for some $k \geq 4$ then the k^{th} derivative of the logged density is bounded. Note also, if the state space was restricted to a compact domain then this condition would be entirely unnecessary anyhow.

As in the setup of the algorithm, assume that there are $n+1$ d-dimensional chains, $\mathbf{x}_0, \ldots, \mathbf{x}_n$, running in parallel at inverse temperature levels, $1 = \beta_0 < \beta_1 <$

$\ldots < \beta_n$ targeting the product distribution

$$\pi_d(\mathbf{x}_0, \ldots, \mathbf{x}_n) \propto f_d^{\beta_0}(\mathbf{x}_0) \ldots f_d^{\beta_n}(\mathbf{x}_n). \tag{3.6}$$

The aim is to determine the optimal spacing between the consecutive temperature levels by considering a pair of neighbouring levels and maximizing the $ESJD_\beta$ given in equation (3.1).

In what follows assume that invariance has been reached and suppose that the algorithm proposes a swap move between the locations of the particles, $\mathbf{x}$ and $\mathbf{y}$, that are at the inverse temperature levels β and $\beta' = \beta + \epsilon$ respectively. Also, suppose that ϵ has the form

$$\epsilon = \frac{\ell}{d^{\frac{1}{2}}} \tag{3.7}$$

where ℓ will be the optimising constant and the particular scaling in d is needed to get a non-trivial asymptotic for the $ESJD_\beta$.

For

$$\mathbf{x} \sim f_d^{\beta} \quad \text{and} \quad \mathbf{y} \sim f_d^{\beta'} \tag{3.8}$$

when the swap move is proposed the locations of the respective particles is adjusted according to the transformation move. Specifically

$$\mathbf{x}' = g(\mathbf{x}, \beta, \beta', \boldsymbol{\mu_x}) = \left(\frac{\beta}{\beta'}\right)^{\frac{1}{2}} (\mathbf{x} - \boldsymbol{\mu_x}) + \boldsymbol{\mu_x} =: g_1(\mathbf{x}) \tag{3.9}$$

$$\mathbf{y}' = g(\mathbf{y}, \beta', \beta, \boldsymbol{\mu_y}) = \left(\frac{\beta'}{\beta}\right)^{\frac{1}{2}} (\mathbf{y} - \boldsymbol{\mu_y}) + \boldsymbol{\mu_y} =: g_2(\mathbf{y}). \tag{3.10}$$

where $\boldsymbol{\mu_z}$ denotes the mode point of the local mode to particle $\mathbf{z}$.

For the marginal targets, f, there is an assumed global maximum at $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ and for the sake of tractability due to issues arising in Chapter 4, let $\boldsymbol{\mu_y} = \boldsymbol{\mu_x} = \boldsymbol{\mu}$. This is a significant and strong assumption. The problem is that the allocation to a mode point essentially partitions the state space into regions, and the mass in each region can be dramatically inconsistent between consecutive temperature levels, see Chapter 4. This would result in a degenerate limit to the $ESJD_\beta$ in this setting.

When the modes are all symmetric in form then the regional mass is preserved. In a symmetric modal setting, without loss of generality (just consider a relocation shift in the state space), the two chains can be considered to be acting in a single mode with centring point $\boldsymbol{\mu}$. The symmetrically multi-modal setting is canonical for the QuanTA algorithm (which was demonstrated to struggle outside

of this context).

Note μ can be any interior point in the state space for the proof to work, not necessarily the mode, but this is the sensible choice.

Then the acceptance probability of the swap move proposed is $\alpha_\beta(\mathbf{x}, \mathbf{y})$ where

$$\alpha_\beta(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{f_d^{\beta'}(g_1(\mathbf{x}))f_d^{\beta}(g_2(\mathbf{y}))}{f_d^{\beta'}(\mathbf{y})f_d^{\beta}(\mathbf{x})}, \tag{3.11}$$

and so recalling equation (1.17) in Section 1.5.1

$$\begin{aligned} ESJD_\beta &= \epsilon^2 \mathbb{E}_{\pi_d}\left[\alpha_\beta(\mathbf{x}, \mathbf{y})\right] \\ &= \epsilon^2 \mathbb{E}_{\pi_d}\left[1 \wedge e^B\right], \end{aligned} \tag{3.12}$$

where

$$B = \log\left(\frac{f_d^{\beta'}(g_1(\mathbf{x}))f_d^{\beta}(g_2(\mathbf{y}))}{f_d^{\beta'}(\mathbf{y})f_d^{\beta}(\mathbf{x})}\right). \tag{3.13}$$

Under the above conditions, the following optimal scaling result will be proved (where $\Phi_{(0,1)}$ is the cumulative distribution function of a standard Normal distribution).

Theorem 3.2.1 (Optimal Scaling for the QuanTA Algorithm)**.** *Consider the parallel tempering algorithm targeting the distribution π_d given in equation* (3.6) *where the target distribution at the cold state ($\beta = 1$) is given by the iid form in equation* (3.2)*. In addition the marginal components of the target are assumed to be regularly varying, satisfying equations* (3.3) *and* (3.4)*, and satisfy the fourth order derivative bound in* (3.5)*. Assuming ϵ scales with dimension as in equation* (3.7) *then as $d \to \infty$, the $ESJD_\beta$, given in equation* (3.12)*, is maximised when ℓ is chosen to maximise*

$$ESJD_\beta = \frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}{\sqrt{2}}\right), \tag{3.14}$$

where

$$\begin{aligned} V(\beta) &= \mathrm{Cov}_\beta((\log f)(x), (x-\mu)(\log f)'(x)) = \frac{1}{\beta^2} \\ I(\beta) &= \mathrm{Var}_\beta\left[(\log f)(x)\right] \\ R(\beta) &= \mathbb{E}_\beta\left[(x-\mu)^2(\log f)''(x) - (x-\mu)(\log f)'(x)\right]. \end{aligned}$$

Furthermore, for the optimal ℓ the corresponding swap move acceptance rate induced between two consecutive temperatures is given by 0.234 (3.s.f).

The proof of this result is given in the following Section 3.3 and is broken down into 3 key steps: computing the appropriate Taylor expansions of the logged swap move acceptance ratio; establishing limiting Gaussianity of the logged swap move acceptance ratio; and finally optimisation of the limiting $ESJD_\beta$.

3.3 Proof of Theorem 3.2.1

3.3.1 Step 1: Taylor Expansion of the Log-Acceptance Ratio

With B as in equation (3.13). Denoting $h(x) := \log(f(x))$ and x_i and y_i to be the i^{th} elements of $\mathbf{x}$ and $\mathbf{y}$ respectively then

$$\begin{aligned} B &= \sum_{i=1}^{d}[\beta' h(g_1(x_i)) - \beta h(x_i)] + \sum_{i=1}^{d}[\beta h(g_2(y_i)) - \beta' h(y_i)] \\ &=: H_\beta^{\beta'}(\mathbf{x}) + H_{\beta'}^{\beta}(\mathbf{y}). \end{aligned} \tag{3.15}$$

With the aim being to derive the asymptotic behaviour of the log acceptance ratio then the next step is to use Taylor expansions (in ϵ) to appropriate order so that the asymptotic behaviour of B can be understood.

For notational convenience, the following will be used:

- Making $h(g_1(x))$ explicitly dependent on ϵ

$$\alpha_x(\epsilon) := h(g_1(x)) = \log\left[f\left(\left(\frac{\beta}{\beta+\epsilon}\right)^{1/2}(x-\mu)+\mu\right)\right].$$

- Denote

$$d_x(\epsilon) := \left(\frac{\beta}{\beta+\epsilon}\right)^{1/2}(x-\mu)+\mu.$$

By Taylor series expansion in ϵ, for fixed x, with Taylor remainder correction term denoted by ξ_x such that $0 < \xi_x < \epsilon$:

$$h(g_1(x)) = \alpha_x(\epsilon) = \alpha_x(0) + \epsilon\alpha_x'(0) + \frac{\epsilon^2}{2}\alpha_x''(0) + \frac{\epsilon^3}{6}\alpha_x'''(\xi_x), \tag{3.16}$$

where

$$\alpha'_x(\epsilon) = -\frac{(x-\mu)}{2}\frac{\beta^{1/2}}{(\beta+\epsilon)^{3/2}}(\log f)'(d_x(\epsilon)), \tag{3.17}$$

$$\begin{aligned}\alpha''_x(\epsilon) &= \frac{(x-\mu)^2}{4}\frac{\beta}{(\beta+\epsilon)^3}(\log f)''(d_x(\epsilon)) \\ &\quad + \frac{3(x-\mu)}{4}\frac{\beta^{1/2}}{(\beta+\epsilon)^{5/2}}(\log f)'(d_x(\epsilon)),\end{aligned} \tag{3.18}$$

$$\begin{aligned}\alpha'''_x(\epsilon) &= -\frac{(x-\mu)^3}{8}\frac{\beta^{3/2}}{(\beta+\epsilon)^{9/2}}(\log f)'''(d_x(\epsilon)) - \frac{9(x-\mu)^2}{8}\frac{\beta}{(\beta+\epsilon)^4}(\log f)''(d_x(\epsilon)) \\ &\quad - \frac{15(x-\mu)}{8}\frac{\beta^{1/2}}{(\beta+\epsilon)^{7/2}}(\log f)'(d_x(\epsilon)).\end{aligned} \tag{3.19}$$

As a preview to the later stages of this proof, the terms up to second order in ϵ dictate the asymptotic distribution of B. However, to show that the higher order terms "disappear" in the limit as $\epsilon \to 0$ then a careful analysis is required. Thus the next step is to establish that, under the assumptions made above, the higher order terms converge to zero in probability.

To this end, a careful analysis of $\alpha'''_x(\cdot)$ is undertaken. Firstly, it will be shown that $|\mathbb{E}_\beta[\alpha'''_x(\xi_x)]|$ is bounded; then application of Markov's inequality will establish that the higher order terms converge to zero in probability as $d \to \infty$. Define

$$\eta_\epsilon := \left[\left(\frac{\beta}{\beta+\epsilon}\right)^{\frac{1}{2}} - 1\right]$$

so that

$$d_x(\epsilon) - x = \left[\left(\frac{\beta}{\beta+\epsilon}\right)^{\frac{1}{2}} - 1\right](x-\mu) := \eta_\epsilon(x-\mu),$$

which has the property that $\eta_\epsilon \to 0$ as $d \to \infty$ and $|\eta_\epsilon| \leq 1$.

Then, with Taylor remainder correction terms denoted $\xi_1^\epsilon, \xi_2^\epsilon, \xi_3^\epsilon$ such that $0 < |\xi_k^\epsilon - x| < |d_x(\epsilon) - x|$

$$\begin{aligned}(\log f)'(d_x(\epsilon)) &= (\log f)'(x) + \eta_\epsilon(x-\mu)(\log f)''(x) + \frac{{\eta_\epsilon}^2(x-\mu)^2}{2}(\log f)'''(x) \\ &\quad + \frac{{\eta_\epsilon}^3(x-\mu)^3}{6}(\log f)''''(\xi_1^\epsilon),\end{aligned} \tag{3.20}$$

$$\begin{aligned}(\log f)''(d_x(\epsilon)) &= (\log f)''(x) + \eta_\epsilon(x-\mu)(\log f)'''(x) \\ &\quad + \frac{{\eta_\epsilon}^2(x-\mu)^2}{2}(\log f)''''(\xi_2^\epsilon),\end{aligned} \tag{3.21}$$

$$(\log f)'''(d_x(\epsilon)) = (\log f)'''(x) + \eta_\epsilon(x-\mu)(\log f)''''(\xi_3^\epsilon). \tag{3.22}$$

Recall the assumptions from equation (3.4) and, in particular, the assumption of boundedness of the fourth order derivatives of $|(\log f)''''(\cdot)| < M$ given in equation (3.5).

Substituting equations (3.20), (3.21) and (3.22) into equation (3.19); evaluating the expectation with respect to $X \sim f^{\beta}$ then $\exists\, C \in \mathbb{R}_+$

$$
\begin{aligned}
& \left|\mathbb{E}_\beta[\alpha_x'''(\xi_x)]\right| \\
\leq\ & \mathbb{E}_\beta\left[\left|\alpha_x'''(\xi_x)\right|\right] \\
\leq\ & \mathbb{E}_\beta\Bigg[\frac{|(x-\mu)|^3}{8}\beta^{-3}|(\log f)'''(d(\xi_x))| + \frac{9|(x-\mu)|^2}{8}\beta^{-3}|(\log f)''(d(\xi_x))| \\
& + \frac{15|(x-\mu)|}{8}\beta^{-3}|(\log f)'(d(\xi_x))|\Bigg] \\
\leq\ & \mathbb{E}_\beta\Bigg[\frac{|(x-\mu)|^3}{8}\beta^{-3}\left(|(\log f)'''(x)| + |(x-\mu)||(\log f)''''(\xi_3^{\xi_x})|\right) \\
& + \frac{9|(x-\mu)|^2}{8}\beta^{-3}\left(|(\log f)''(x)| + |(x-\mu)||(\log f)'''(x)| + \frac{|x|^2}{2}|(\log f)''''(\xi_2^{\xi_x})|\right) \\
& + \frac{15|(x-\mu)|}{8}\beta^{-3}\Bigg(|(\log f)'(x)| + |(x-\mu)||(\log f)''(x)| + \frac{|(x-\mu)|^2}{2}|(\log f)'''(x)| \\
& + \frac{|(x-\mu)|^3}{6}|(\log f)''''(\xi_1^{\xi_x})|\Bigg)\Bigg] \\
\leq\ & C
\end{aligned}
\tag{3.23}
$$

where the first three inequalities are from the direct application of the triangle inequality (with the second also using the boundedness of η_ϵ); whereas the final inequality arises from both the finiteness of expectations of the terms involving derivatives of order three or below (this is due to the regularly varying tails of $\log(f(\cdot))$) and the assumption that $|(\log f)''''(\cdot)| < M$.

Using equation (3.16), with substitution of terms from equations (3.17), (3.18) and (3.19), $H_\beta^{\beta'}(\mathbf{x})$ can be expressed as

$$
\begin{aligned}
H_\beta^{\beta'}(\mathbf{x}) &= \sum_{i=1}^{d}(\beta+\epsilon)\left[\alpha_{x_i}(\epsilon) - \beta\alpha_{x_i}(0)\right] \\
&= \epsilon\sum_{i=1}^{d}\left[\alpha_{x_i}(0) + \beta\alpha_{x_i}'(0)\right] + \epsilon^2\sum_{i=1}^{d}\left[\frac{\beta}{2}\alpha_{x_i}''(0) + \alpha_{x_i}'(0)\right] \\
&\quad + \epsilon^3\sum_{i=1}^{d}\left[\frac{1}{2}\alpha_{x_i}''(0) + \frac{\beta}{6}\alpha_{x_i}'''(\xi_{x_i})\right] + \epsilon^4\sum_{i=1}^{d}\frac{1}{6}\alpha_{x_i}'''(\xi_{x_i}).
\end{aligned}
\tag{3.24}
$$

By equation (3.23) and using the iid nature of the $x_i's$ and using Markov's inequality then $\forall \delta > 0$

$$\begin{aligned} & \delta \mathbb{P}\left(\left| \epsilon^3 \sum_{i=1}^{d} \left[\frac{1}{2}\alpha''_{x_i}(0) + \frac{\beta}{6}\alpha'''_{x_i}(\xi_{x_i}) \right] \right| > \delta \right) \\ < & \mathbb{E}\left(\left| \frac{\ell^3}{d^{3/2}} \sum_{i=1}^{d} \left[\frac{1}{2}\alpha''_{x_i}(0) + \frac{\beta}{6}\alpha'''_{x_i}(\xi_{x_i}) \right] \right| \right) \\ \leq & \frac{\ell^3}{d^{1/2}} \left[\frac{1}{2}\mathbb{E}\left(|\alpha''_{x_i}(0)| \right) + \frac{\beta}{6}C \right] \to 0 \ \text{ as } \ d \to \infty. \end{aligned}$$

Thus,

$$\epsilon^3 \sum_{i=1}^{d} \left[\frac{1}{2}\alpha''_{x_i}(0) + \frac{\beta}{6}\alpha'''_{x_i}(\xi_{x_i}) \right] \to 0 \ \text{ in probability as } \ d \to \infty.$$

By identical methodology, as $d \to \infty$

$$\epsilon^4 \sum_{i=1}^{d} \frac{1}{6}\alpha'''_{x_i}(\xi_{x_i}) \to 0 \ \text{ in probability.}$$

Consequently,

$$\begin{aligned} H_\beta^{\beta'}(\mathbf{x}) &= \epsilon \left[\sum_{i=1}^{d} h(x_i) - \frac{1}{2}(x_i - \mu)h'(x_i) \right] \\ &\quad + \frac{\epsilon^2}{8\beta} \left[\sum_{i=1}^{d} (x_i - \mu)^2 h''(x_i) - (x_i - \mu)h'(x_i) \right] + T_x \end{aligned} \tag{3.25}$$

where

$$T_x = \epsilon^3 \sum_{i=1}^{d} \left[\frac{1}{2}\alpha''_{x_i}(0) + \frac{\beta}{6}\alpha'''_{x_i}(\xi_{x_i}) \right] + \epsilon^4 \sum_{i=1}^{d} \frac{1}{6}\alpha'''_{x_i}(\xi_{x_i})$$

with $T_x \to 0$ in probability as $d \to \infty$.

Now denoting $h(g_2(y))$ as

$$\alpha_y(\epsilon) := h(g_2(y)) = \log \left[f\left(\left(\frac{\beta + \epsilon}{\beta} \right)^{1/2} (y - \mu) + \mu \right) \right],$$

the Taylor series expansion in ϵ, for a fixed y, with Taylor truncation term denoted by ξ_y such that $0 < \xi_y < \epsilon$ is given by

$$h(g_2(x)) = \alpha_y(\epsilon) = \alpha_y(0) + \epsilon\alpha'_y(0) + \frac{\epsilon^2}{2}\alpha''_y(0) + \frac{\epsilon^3}{6}\alpha_y(\xi_y). \tag{3.26}$$

By identical methodology to the above calculation in equation (3.23) for $\alpha_x(\cdot)$, it can be shown that $\exists\ C_y \in \mathbb{R}_+$ such that

$$\left|\mathbb{E}_\beta[\alpha_y'''(\xi_y)]\right| \ \leq \ C_y. \tag{3.27}$$

Hence, using exactly the same methodology as for the x_i's above, then

$$\begin{aligned} H^\beta_{\beta'}(\mathbf{y}) \ &= \ -\epsilon\left[\sum_{i=1}^d h(y_i) - \frac{1}{2}(y_i-\mu)h'(y_i)\right] \\ &\quad + \frac{\epsilon^2}{8\beta}\left[\sum_{i=1}^d (y_i-\mu)^2 h''(y_i) - (y_i-\mu)h'(y_i)\right] + T_y. \end{aligned} \tag{3.28}$$

where $T_y \to 0$ in probability as $d \to \infty$.

Two definitions will now be established that will be useful for notational convenience but also highlight important terms that will appear in the final result.

Definition 3.3.1. With notation established thus far in the statement and proof of Theorem 3.2.1 then the following are defined:

- $k(z) = (z-\mu)h'(z)$
- $r(z) = (z-\mu)^2 h''(z)$

Using this new notation the log acceptance ratio, B, from equation (3.15) is written as

$$\begin{aligned} B \ &= \ \epsilon\left[\sum_{i=1}^d h(x_i) - h(y_i) + \frac{1}{2}\left(k(y_i) - k(x_i)\right)\right] \\ &\quad + \frac{\epsilon^2}{8\beta}\left[\sum_{i=1}^d r(x_i) - k(x_i) + r(y_i) - k(y_i)\right] + (T_x + T_y). \end{aligned} \tag{3.29}$$

Next some notation is introduced that will be useful for the following steps of the argument. Define

$$M(\beta) \ = \ \mathbb{E}_\beta(h(z)) \tag{3.30}$$

$$S(\beta) \ = \ \mathbb{E}_\beta(k(z)) \tag{3.31}$$

$$R(\beta) \ = \ \mathbb{E}_\beta(r(z) - k(z)), \tag{3.32}$$

where the $M(\beta)$ is as it was in Atchadé *et al.* [2011]. In all cases the expectation is with respect to the distribution $\frac{f^\beta(x)}{Z_\beta}$ where $Z_\beta = \int f^\beta(z)dz$. By Taylor expansion

to first order then $M(\beta+\epsilon) - M(\beta) = \epsilon M'(\beta) + O(\epsilon^2)$ where

$$\begin{aligned} I(\beta) := M'(\beta) &= \int (\log f(x))^2 \frac{f^\beta(x)}{Z_\beta} dx - \left[\int \log f(x) \frac{f^\beta(x)}{Z_\beta} dx\right]^2 \\ &= \mathbb{E}_\beta(h(x)^2) - [\mathbb{E}_\beta(h(x))]^2 \\ &= \mathrm{Var}_\beta(h(x)). \end{aligned} \tag{3.33}$$

An important identity for the term $S(\beta)$ is also given by

$$\begin{aligned} S(\beta) &= \int (x-\mu)(\log f)'(x) \frac{f^\beta(x)}{Z_\beta} dx \\ &= \int (x-\mu) f'(x) \frac{f^{\beta-1}(x)}{Z_\beta} dx \\ &= \cancel{\left[\frac{(x-\mu)\, f^\beta(x)}{\beta\, Z_\beta}\right]_{-\infty}^{-\infty}} \overset{0}{} - \int \frac{1}{\beta} \frac{f^\beta(x)}{Z_\beta} dx = -\frac{1}{\beta}, \end{aligned} \tag{3.34}$$

where the last step follows as the expectation of the score function is zero. Again by Taylor expansion to first order $S(\beta+\epsilon) - S(\beta) = \epsilon S'(\beta) + O(\epsilon^2)$. From the form of $S(\beta)$ in equation (3.34) it is clear that for all f,

$$V(\beta) := S'(\beta) = \frac{1}{\beta^2} \tag{3.35}$$

By taking the integral form of $S(\beta)$ and differentiating with respect to β, similar to (3.33), it can be shown that

$$V(\beta) = \mathbb{E}_\beta[k(x)h(x)] - \mathbb{E}_\beta[k(x)]\mathbb{E}_\beta[h(x)] = \mathrm{Cov}_\beta(h(x), k(x)). \tag{3.36}$$

This is simply a nice observation; it is the form of $V(\beta)$ given in equation (3.35) that will be useful herein.

3.3.2 Step 2: Establising the Asymptotic Normality of B

It will now be established that B (in equation (3.13)) is asymptotically Gaussian of the form $N(-c, 2c)$, for some c, as $d \to \infty$ provided $\epsilon = \ell/d^{1/2}$.

Now, making the dimensionality dependence explicit, write $B = W(d) +$

$(T_x + T_y)$ where

$$\begin{aligned} W(d) &:= \epsilon \left[\sum_{i=1}^{d} h(x_i) - h(y_i) + \frac{1}{2} \left(k(y_j) - k(x_j) \right) \right] \\ &\quad + \frac{\epsilon^2}{8\beta} \left[\sum_{i=1}^{d} r(x_i) - k(x_i) + r(y_i) - k(y_i) \right] \end{aligned}$$

and $(T_x + T_y) \to 0$ in probability as $d \to \infty$. Hence, if it can be shown that $W(d)$ converges in distribution to a Gaussian of the form $N(-c, 2c)$ then by Slutsky's Theorem one can conclude that B converges in distribution to the same Gaussian as the W.

To this end, the asymptotic Gaussianity of $W(d)$ is established. First note that due to the iid nature of the x_i's and y_i's respectively then by the standard Central Limit Theorem, e.g. Durrett [2010], for a sum of iid variables, then asymptotic Gaussianity is immediate where

$$W(d) \Rightarrow N\left(\mu_W, \sigma_W^2\right) \quad \text{as} \quad d \to \infty \tag{3.37}$$

where

$$\mu_W = \lim_{d \to \infty} \mathbb{E}[W(d)]$$

and

$$\sigma_W^2 = \lim_{d \to \infty} \mathrm{Var}[W(d)].$$

To this end the terms $\mathbb{E}[W(d)]$ and $\mathrm{Var}[W(d)]$ are computed.

$$\begin{aligned} \mathbb{E}[W(d)] &:= \epsilon \left[\sum_{i=1}^{d} M(\beta) - M(\beta + \epsilon) - \frac{1}{2}(S(\beta) - S(\beta + \epsilon)) \right] \\ &\quad + \frac{\epsilon^2}{8\beta} \left[\sum_{i=1}^{d} R(\beta) + R(\beta + \epsilon) \right] \\ &= \epsilon \left[\sum_{i=1}^{d} -\epsilon M'(\beta) + \frac{\epsilon}{2} S'(\beta) \right] + \frac{\epsilon^2}{8\beta} \left[\sum_{i=1}^{d} 2R(\beta) \right] + O(d^{-1/2}) \\ &\quad \to \ell^2 \left[\frac{1}{2} V(\beta) - I(\beta) + \frac{1}{4\beta} R(\beta) \right] \quad \text{as} \quad d \to \infty. \end{aligned}$$

Similarly,

$$\mathrm{Var}(W(d)) \to 2\ell^2 \mathrm{Var}_\beta \left(h(x) - \frac{1}{2} k(x) \right) \quad \text{as} \quad d \to \infty.$$

Hence by Slutsky's Theorem then B is asymptotically Gaussian such that

$$B \;\dot{\sim}\; N\left(\ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right], 2\ell^2 \mathrm{Var}_\beta\left(h(x) - \frac{1}{2}k(x)\right)\right). \tag{3.38}$$

Note that for any general target π and proposal q then the Metropolis-Hastings acceptance ratio has the property that

$$\mathbb{E}_{\pi,q}\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) = \int\int \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\pi(x)q(x,y)dydx = 1 \tag{3.39}$$

and so it is key that if the above expressions for B is correct and the right asymptotic form for B has been found then $\lim_{d\to\infty}\mathbb{E}\left(e^B\right) = 1$ and thus one requires

$$\mu_W = -\frac{\sigma_W^2}{2}. \tag{3.40}$$

At first it is not obvious that the limiting Gaussian derived for B has this essential property. For the sake of completeness this will be verified before moving on to optimisation of the temperature level spacings.

Lemma 3.3.1. *With notations established in Chapter 3 and in particular under the assumptions of Theorem 3.2.1 then*

$$\ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = -\ell^2\mathrm{Var}_\beta\left(h(x) - \frac{1}{2}k(x)\right), \tag{3.41}$$

which ensures that the key identity in equation (3.40) *is satisfied for the limiting form for B in equation* (3.38)*.*

Proof. From equation (3.38) then denote

$$\mu = \ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] \tag{3.42}$$

and

$$\sigma^2 = 2\ell^2\mathrm{Var}_\beta\left(h(x) - \frac{1}{2}k(x)\right).$$

Then by using the standard properties of variance it is routine to show that

$$-\frac{\sigma^2}{2} = \ell^2\left[-I(\beta) - \frac{1}{4}\mathrm{Var}_\beta(k(x)) + V(\beta)\right]. \tag{3.43}$$

Consequently, equating the terms on the RHS of equations (3.42) and (3.43) shows that if the following can be shown to hold then the required identity in equa-

tion (3.41) is validated:

$$\frac{1}{4\beta}R(\beta) = -\frac{1}{4}\text{var}_\beta(k(x)) + \frac{1}{2}V(\beta). \tag{3.44}$$

The LHS and RHS of equation (3.44) will be considered separately. The following integration by parts are well defined due to the assumption that $-\log(f(\cdot))$ has regularly varying tails. Starting with the RHS and recalling that from equation (3.34) $\mathbb{E}_\beta(k(x)) = -1/\beta$:

$$\begin{aligned} -\frac{1}{4}\text{var}_\beta(k(x)) + \frac{1}{2}V(\beta) &= -\frac{1}{4}\left[\mathbb{E}_\beta(k(x)^2) - \mathbb{E}_\beta(k(x))^2\right] + \frac{1}{2\beta^2} \\ &= -\frac{1}{4}\mathbb{E}_\beta(k(x)^2) + \frac{3}{4\beta^2}. \end{aligned}$$

Then, noting that $(\log f)'(x)f^\beta(x) = f'(x)f^{\beta-1}(x)$, and using integration by parts (by first integrating $f'(x)f^{\beta-1}(x)$):

$$\begin{aligned} \mathbb{E}_\beta(k(x)^2) &= \int (x-\mu)^2[(\log f)'(x)]^2 \frac{f^\beta(x)}{Z_\beta}dx \\ &= \cancel{\left[\frac{(x-\mu)^2}{\beta}(\log f)'(x)\frac{f^\beta(x)}{Z_\beta}\right]_{-\infty}^{-\infty}} \;\to 0 \\ &\quad -\frac{1}{\beta}\int \left[(x-\mu)^2(\log f)''(x) + 2(x-\mu)(\log f)'(x)\right]\frac{f^\beta(x)}{Z_\beta}dx \\ &= -\frac{1}{\beta}\mathbb{E}_\beta(r(x)) - \frac{2}{\beta}\mathbb{E}_\beta(k(x)) = -\frac{1}{\beta}\mathbb{E}_\beta(r(x)) + \frac{2}{\beta^2}. \end{aligned} \tag{3.45}$$

Collating the above in equations (3.44) and (3.45) then

$$-\frac{1}{4}\text{var}_\beta(k(x)) + \frac{1}{2}V(\beta) = \frac{1}{4\beta}\mathbb{E}_\beta(r(x)) + \frac{1}{4\beta^2} = \frac{1}{4\beta}R(\beta), \tag{3.46}$$

where the final equality simply comes from the definition of $R(\beta)$ from equation (3.32). □

Using Lemma 3.3.1 it is concluded that $B \dot{\sim} N(\frac{-\sigma^2}{2}, \sigma^2)$ where

$$\sigma^2 = 2\ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right].$$

3.3.3 Step 3: Optimisation

Firstly, an auxiliary result, given in Roberts *et al.* [1997], that will help in the process of deriving the optimal spacings will be established in the following calculation.

Letting $\phi_{(m,\sigma^2)}$ denote the density function of a Gaussian with mean m and variance σ^2 and suppose that $G \sim N(-\frac{\sigma^2}{2}, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}(1 \wedge e^G) &= \int_0^\infty \phi_{\left(-\frac{\sigma^2}{2},\sigma^2\right)}(g)dg + \int_{-\infty}^0 e^g \phi_{\left(-\frac{\sigma^2}{2},\sigma^2\right)}(g)dg \\
&= 1 - \Phi_{\left(-\frac{\sigma^2}{2},\sigma^2\right)}(0) + \Phi_{\left(\frac{\sigma^2}{2},\sigma^2\right)}(0) \\
&= 1 - \Phi_{(0,1)}\left(\frac{\sigma}{2}\right) + \Phi_{(0,1)}\left(-\frac{\sigma}{2}\right) \\
&= 2\Phi_{(0,1)}\left(-\frac{\sigma}{2}\right). \qquad (3.47)
\end{aligned}$$

Using the result in equation (3.47), for large d, and recalling the form of the $ESJD_\beta$ from equation (3.12) then

$$ESJD_\beta = \frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}{\sqrt{2}}\right). \qquad (3.48)$$

Letting

$$u = \ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}, \qquad (3.49)$$

then

$$ESJD_\beta = \frac{2u^2}{d\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]}\Phi_{(0,1)}\left(-\frac{u}{\sqrt{2}}\right). \qquad (3.50)$$

Consider optimising the $ESJD_\beta$ now with respect to u. It is clear that the optimising value, denoted u^*, doesn't depend on

$$\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right].$$

Determining u^*'s value explicitly can be done numerically.

Recalling the form of $ESJD_\beta$ in equation (3.12) then for every choice of u there is an associated acceptance rate, denoted here by ACC_β. Indeed, for the optimising value u^* there is an associated optimal acceptance rate

$$\mathrm{ACC}^*_\beta = 2\Phi_{(0,1)}\left(-\frac{u^*}{\sqrt{2}}\right). \qquad (3.51)$$

Indeed, if u^* is the maximiser of equation (3.50) then

$$\text{ACC}_\beta^* = 0.234 \quad (3.s.f) \tag{3.52}$$

and this consequently completes the proof of Theorem 3.2.1.

3.4 Interpretation and Discussion of Theorem 3.2.1

An important sanity check for the result in Theorem 3.2.1 is to interpret the result when the target is in the most basic version of the canonical setting. This is when the marginal targets are Gaussian with mean μ and standard deviation σ, and so $f(x) \propto \phi_{(\mu,\sigma^2)}(x)$.

It has already been illustrated in Section 2.2 that in this setting a temperature swap move with any spacing can be made with an acceptance probability of one. In the setting of optimal spacing this implies that there is no limit to the ambitiousness of the spacing. Corollary 3.4.1 shows this is indeed the case for the Gaussian setting which is reasuring.

Corollary 3.4.1. *Under the setting of Theorem 3.2.1 where $f(x) \propto \phi_{(\mu,\sigma^2)}(x)$ then the asymptotic $ESJD_\beta$ takes the form*

$$ESJD_\beta = \frac{2\ell^2}{d}\Phi_{(0,1)}(0) \tag{3.53}$$

which shows that there is no finite value of ℓ that maximises the $ESJD_\beta$ and ℓ can be chosen arbitrarily large irrespective of the dimension.

Proof. Equation (3.14) from the statement of Theorem 3.2.1 gives the general form of the $ESJD_\beta$ for a general marginal f. If it can be shown that for $f(x) \propto \phi_{(\mu,\sigma^2)}(x)$

$$\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = 0 \tag{3.54}$$

then by appealing to the form of the $ESJD_\beta$ given in equation (3.48) it is clear that there is no finite ℓ that maximises the $ESJD_\beta$ and so the Corollary is proved.

Without loss of generality, assume that each marginal is a standard normal with 0 mean and variance 1. Then

- $V(\beta) = \frac{1}{\beta^2}$
- $I(\beta) = \text{Var}_\beta(-\frac{x^2}{2}) = \frac{1}{2\beta^2}$

- $R(\beta) = \mathbb{E}_\beta\left(-\frac{x^2}{2} + \frac{x^2}{2}\right) = 0.$

Combining these three identities as in the LHS of equation (3.54) then it is clear that this equates to 0 as required. □

3.4.1 Higher Order Scalings at Cold Temperatures

For this section let $\phi(\cdot)$ denote the density function of the standard Gaussian distribution.

Recall that $I(\beta) = 1/(2\beta^2)$ for any uni-variate Gaussian distribution at inverse temperature level β. Section 1.5.2 showed that the optimal spacings result of Atchadé *et al.* [2011] implies that the optimal choice for the scaling parameter takes the form

$$\hat{\ell} \propto \beta \tag{3.55}$$

resulting in a geometrically spaced temperature schedule.

Assuming appropriate smoothness for the iid marginal target densities, $f(\cdot)$'s, then for a sufficiently cold temperature the local mode can be well approximated by a Gaussian. Hence for sufficiently cold temperatures then one expects to see that $I(\beta) \approx 1/(2\beta^2)$; thus inducing an (approximately) geometrically spaced temperature schedule at these cold temperature levels. A rigorous derivation that $I(\beta) \approx 1/(2\beta^2)$ is contained in the proof of Theorem 3.4.1 below.

For the QuanTA algorithm the optimal spacing in the canonical Gaussian case is of "infinitely" higher order with regards to the temperature level since there is no restriction on the size of the temperature spacings with regards the value of β.

In the case where the target distribution is sufficiently smooth then $\pi_\beta(x)$ converges to a Gaussian as $\beta \to \infty$. This suggests that even outside of the Gaussian setting, once in the super colder temperatures, the QuanTA approach will exhibit a higher order behaviour (with respect to β) for the spacings. Equation (3.55) showed that the spacings for a Gaussian target in a standard PT setting are $O(\beta)$. So for cold temperatures one should expect QuanTA to permit higher order behaviour so that the spacings would be $O(\beta^\zeta)$ where $\zeta > 1$. The following Theorem 3.4.1 will establish when this is the case.

However, before the statement of the theorem, three assumption statements are made about the target density $f(\cdot)$, as given in Theorem 3.2.1. These provide a sufficient form of $f(\cdot)$ for proving the theorem.

Assume that the marginal component of the target, $f(\cdot)$, is uni-modal with mode point at $\mu = 0$ without loss of generality. Also, assume that $f(\cdot)$ is in C^4 and

define the normalised density $g_\beta(\cdot)$ such that

$$g_\beta(y) \propto f^\beta\left(\mu + \frac{y}{\sqrt{-\beta(\log f)''(\mu)}}\right) = f^\beta\left(\frac{y}{\sqrt{-\beta(\log f)''(0)}}\right) \tag{3.56}$$

1. For $\gamma > 0$ as $\beta \to \infty$

$$\left|\mathrm{Var}_{g_\beta}\left(Y^2\right) - 2\right| = O\left(\frac{1}{\beta^\gamma}\right); \tag{3.57}$$

2. Bounded fourth derivatives of $\log f(\cdot)$ i.e. there exists a positive constant M such that for all $z \in \mathbb{R}$

$$\left|(\log f)''''(z)\right| < M; \tag{3.58}$$

3. For all $\beta > 0$, the eighth moment exists and is finite

$$\mathbb{E}_\beta\left[X^8\right] = \int_{\mathcal{X}} x^8 \frac{\pi^\beta(x)}{\int_{\mathcal{X}} \pi^\beta(z)dz} dx < \infty. \tag{3.59}$$

Note that in Theorem 3.4.1, then the conditions on $f(\cdot)$ are inherited from the conditions on $f(\cdot)$ from Theorem 3.2.1. Thus assumptions 2 and 3 are implicitly satisfied but are still stated due to their importance for proving the result.

Theorem 3.4.1 (Cold Temperature Scalings)**.** *Under the setting of Theorem 3.2.1 then the optimal $ESJD_\beta$ is derived by maximising*

$$\frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}{\sqrt{2}}\right)$$

with respect to ℓ.

For large β, if the marginal target, $f(\cdot)$ satisfies the three conditions/assumptions in equations (3.57)*,* (3.58) *and* (3.59) *given above, then*

$$\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = O\left(\frac{1}{\beta^k}\right),$$

where

- $k = \min\left\{2 + \gamma, 3\right\} > 2$ *if f is symmetric about the mode point 0*
- $k = \min\left\{2 + \gamma, \frac{5}{2}\right\} > 2$ *otherwise.*

This induces an optimising value $\hat{\ell}$ such that

$$\hat{\ell} = O\left(\beta^{\frac{k}{2}}\right), \tag{3.60}$$

showing that at the colder temperatures QuanTA permits higher order behaviour than the standard PT scheme which has $\hat{\ell} = O(\beta)$.

Remark: For the case that f is not symmetric about the mode point then Assumptions 2 and 3 can be altered respectively so that only third derivatives of $\log f(\cdot)$ need to be bounded and only sixth moments (rather than eighth) need to exist.

Proof. Recall the definition of u from equation (3.49). Denoting by u^* the value of u that maximises $ESJD_\beta$ and it is immediate from equation (3.49) that

$$\hat{\ell} = \frac{u^*}{\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}. \tag{3.61}$$

As a result, if it can be shown that

$$\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = O\left(\frac{1}{\beta^k}\right), \tag{3.62}$$

then the result of the theorem follows. To this end, the LHS of equation (3.62) will be split into two terms which are analysed individually. For notational convenience, $h(\cdot) := \log f(\cdot)$ with corresponding derivatives denoted $h'(\cdot)$, $h''(\cdot)$, $h'''(\cdot)$ etc.

The $\frac{1}{2}V(\beta) - I(\beta)$ term:

It has already been established that $V(\beta) = 1/\beta^2$ for all distributions. Also, for a Gaussian density, $f(\cdot)$, $I(\beta) = 1/(2\beta^2)$. Since $g_\beta(\cdot)$ approaches the density of a standard Gaussian, $\phi(\cdot)$, as $\beta \to \infty$, then one expects that $I(\beta)$ would approach $1/(2\beta^2)$ too. Hence, a rigorous analysis of this convergence needs to be established. Note that

$$\begin{aligned} I(\beta) &= \mathrm{Var}_\beta\left[h(X)\right] \\ &= \int \left(h(x) - \mathbb{E}_\beta[h(X)]\right)^2 \frac{f^\beta(x)}{Z(\beta)} dx \\ &= \int \left(h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) - \mathbb{E}_{g_\beta}\left[h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right]\right)^2 g_\beta(y) dy \end{aligned} \tag{3.63}$$

using the change of variable, $X = \frac{Y}{\sqrt{\beta(-h''(0))}}$. By Taylor expansion of h about the mode point, 0, up to fourth order then

$$\begin{aligned} h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) &= h(0) - \frac{y^2}{2\beta} \\ &\quad + \frac{y^3 h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}} + \frac{y^4 h''''(\xi_1(y))}{24\left(\beta(-h''(0))\right)^2} \end{aligned} \tag{3.64}$$

where $\xi_1(\cdot)$ is the truncation term for the Taylor expansion such that

$$0 < |\xi_1(y)| < \left|\frac{y}{\sqrt{\beta(-h''(0))}}\right|$$

for all y. Using the Taylor expansion form of h and assumption 2 bounding the fourth derivatives given in equation (3.58)

$$\begin{aligned} &\left|\mathbb{E}_{g_\beta}\left[h\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right) - h(0) + \frac{Y^2}{2\beta} - \frac{Y^3 h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}}\right]\right| \\ &\quad \leq \mathbb{E}_{g_\beta}\left[\left|\frac{Y^4 h''''(\xi_1(Y))}{24\left(\beta(-h''(0))\right)^2}\right|\right] \leq \frac{M}{24\left(\beta(-h''(0))\right)^2}\mathbb{E}_{g_\beta}\left[Y^4\right] = O\left(\frac{1}{\beta^2}\right) \end{aligned}$$

where $\mathbb{E}_{g_\beta}\left[Y^4\right] < \infty$ by assumption 3 given in equation (3.59). Thus,

$$\mathbb{E}_{g_\beta}\left[h\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)\right] = h(0) - \frac{\mathbb{E}_{g_\beta}\left[Y^2\right]}{2\beta} + \frac{\mathbb{E}_{g_\beta}\left[Y^3\right]h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}} + \frac{\mathbb{E}_{g_\beta}\left[Y^4 h''''(\xi_1(Y))\right]}{24\left(\beta(-h''(0))\right)^2},$$

and substituting this into equation (3.63), along with the Taylor expansion of h to the fourth order given in equation (3.64), gives

$$\begin{aligned} I(\beta) &= \int \left(h(0) - \frac{y^2}{2\beta} + \frac{y^3 h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}} + \frac{y^4 h''''(\xi_1(y))}{24\left(\beta(-h''(0))\right)^2}\right. \\ &\quad \left. - \left[h(0) + \frac{\mathbb{E}_{g_\beta}\left[Y^2\right]h''(0)}{2\beta(-h''(0))} + \frac{\mathbb{E}_{g_\beta}\left[Y^3\right]h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}} + \frac{\mathbb{E}_{g_\beta}\left[Y^4 h''''(\xi_1(Y))\right]}{24\left(\beta(-h''(0))\right)^2}\right]\right)^2 g_\beta(y)dy \\ &= \frac{1}{4\beta^2}\int \left(y^2 - \mathbb{E}_{g_\beta}\left[Y^2\right]\right)^2 g_\beta(y)dy \\ &\quad + \frac{2h'''(0)}{24\beta^{5/2}(-h''(0))^{3/2}}\int \left(y^2 - \mathbb{E}_{g_\beta}\left[Y^2\right]\right)\left(y^3 - \mathbb{E}_{g_\beta}\left[Y^3\right]\right) g_\beta(y)dy \\ &\quad + O\left(\frac{1}{\beta^3}\right), \end{aligned}$$

which is finite and well defined due to assumptions 2 and 3. Consequently, in general

$$I(\beta) = \frac{1}{4\beta^2}\text{Var}_{g_\beta}\left(Y^2\right) + O\left(\frac{1}{\beta^{5/2}}\right),$$

but in the case that $h'''(0) = 0$, which indeed holds in the case that f is symmetric about the mode point, then

$$I(\beta) = \frac{1}{2\beta^2}\text{Var}_{g_\beta}\left(Y^2\right) + O\left(\frac{1}{\beta^3}\right)$$

and so under assumption 1 given in equation (3.57), then

$$I(\beta) = \frac{1}{2\beta^2} + O\left(\frac{1}{\beta^k}\right) \tag{3.65}$$

where in general $k = \min\{2+\gamma, 5/2\}$ but if $h'''(0) = 0$ then $k = \min\{2+\gamma, 3\}$, and so $\frac{1}{2}V(\beta) - I(\beta) = O\left(\frac{1}{\beta^k}\right)$.

The $R(\beta)$ term:

Recall that

$$\begin{aligned} R(\beta) &= \frac{1}{4\beta}\mathbb{E}_\beta\left[X^2h''(X) - Xh'(X)\right] \\ &= \frac{1}{4\beta}\mathbb{E}_{g_\beta}\left[\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)^2 h''\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)\right. \\ &\qquad \left. -\frac{Y}{\sqrt{\beta(-h''(0))}}h'\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)\right]. \end{aligned} \tag{3.66}$$

Using Taylor expansion about the mode at 0 then

$$\begin{aligned} h'\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) &= h'(0) + \frac{y}{\sqrt{\beta(-h''(0))}}h''(0) + \frac{y^2}{2\beta(-h''(0))}h'''(0) \\ &\quad + \frac{y^3}{6\beta^{3/2}(-h''(0))^{3/2}}h''''(\xi_2(y)), \end{aligned} \tag{3.67}$$

where $\xi_2(\cdot)$ is the truncation term for the Taylor expansion such that

$$0 < |\xi_2(y)| < \left|\frac{y}{\sqrt{\beta(-h''(0))}}\right|$$

for all y. Also,

$$h''\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) = h''(0) + \frac{y}{\sqrt{\beta(-h''(0))}}h'''(0) + \frac{y^2}{2\beta^{3/2}(-h''(0))^{3/2}}h''''(\xi_3(y)) \quad (3.68)$$

where $\xi_3(\cdot)$ is the truncation term for the Taylor expansion such that $0 < |\xi_3(y)| < |y|$ for all y. Hence,

$$\frac{y^2}{2\beta(-h''(0))}h''\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) - \frac{y}{\sqrt{\beta(-h''(0))}}h'\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right)$$
$$= \frac{y^3}{2\left(\beta(-h''(0))\right)^{3/2}}h'''(0) + \frac{y^4}{\left(\beta(-h''(0))\right)^2}\left[\frac{1}{2}h''''(\xi_3(y)) - \frac{1}{6}h''''(\xi_2(y))\right].$$

Substituting this in to the $R(\beta)$ term in equation (3.66)

$$\begin{aligned} R(\beta) &= \frac{1}{4\beta}\mathbb{E}_{g_\beta}\left[\frac{Y^3}{2\left(\beta(-h''(0))\right)^{3/2}}h'''(0) + \frac{Y^4}{\left(\beta(-h''(0))\right)^2}\left[\frac{1}{2}h''''(\xi_3(Y)) - \frac{1}{6}h''''(\xi_2(Y))\right]\right] \\ &= \frac{h'''(0)}{8\beta^{5/2}(-h''(0))^{3/2}}\mathbb{E}_{g_\beta}\left[Y^3\right] + \frac{1}{4\beta^3(-h''(0))^2}\mathbb{E}_{g_\beta}\left[Y^4\left[\frac{1}{2}h''''(\xi_3(Y)) - \frac{1}{6}h''''(\xi_2(Y))\right]\right], \end{aligned}$$

where

$$\mathbb{E}_{g_\beta}\left[Y^4\left[\frac{1}{2}h''''(\xi_3(Y)) - \frac{1}{6}h''''(\xi_2(Y))\right]\right] < \infty$$

due to assumptions 2 and 3 given in (3.58) and (3.59) respectively. Hence, in general

$$R(\beta) = O\left(\frac{1}{\beta^{5/2}}\right)$$

but in the case that $h'''(\cdot) = 0$, which is the case when $f(\cdot)$ is symmetric about the mode point 0, then

$$R(\beta) = O\left(\frac{1}{\beta^3}\right).$$

Consequently,

$$R(\beta) = O\left(\frac{1}{\beta^k}\right) \quad (3.69)$$

where in general $k = 5/2$ but in the case that $h'''(0) = 0$ then $k = 3$.

Combining the results of equations (3.65) and (3.69) completes the proof. □

Theorem 3.4.1 shows that the QuanTA approach gives higher order behaviour

in the limit as $\beta \to \infty$. However, this was under the assumption that

$$|\text{Var}_{g_\beta}\left(Y^2\right) - 2| = O\left(\frac{1}{\beta^\gamma}\right).$$

but one should analyse the realism of this assumption. Heuristically, since $g_\beta(\cdot)$ approaches a Gaussian density then one expects Y^2 to approach a χ^2 random variable on 1 degree of freedom which would have a variance of 2.

To gain insight into typical γ values, next are derived the γ values in the cases that $f(\cdot)$ is the density of a Gamma(a, b) distribution and a t-distribution on ν degrees of freedom respectively. It will be shown in both cases that as $\beta \to \infty$ the respective densities approach a Gaussian and also that the rate of convergence in the assumption is $\gamma = 1$ in both cases.

Gamma Example:

Suppose that $f(\cdot)$ is the density function of a Gamma(a, b) with $a > 1$ and so

$$f(x) \propto x^{a-1} \exp\{-bx\}$$

and so

$$f^\beta(x) \propto x^{\beta(a-1)} \exp\{-\beta bx\}$$

hence with slight abuse of notation $f^\beta \sim \text{Gamma}(\beta(a-1)+1, \beta b)$.

Now by routine calculation it can be shown that the mode point is given by

$$\mu = \frac{a-1}{b}$$

and

$$h''(\mu) = -\frac{b^2}{a-1}.$$

As in the statement of Theorem 3.4.1, but now with non-zero mode point, $g_\beta(\cdot)$ is defined as

$$g_\beta(y) \propto f^\beta\left(\mu + \frac{y}{\sqrt{-\beta h''(\mu)}}\right)$$

which is achieved by transforming the original random variable, $X \sim f^\beta$, such that

$$X = \mu + \frac{Y}{\sqrt{-\beta h''(\mu)}}.$$

Firstly, it will be shown that $g_\beta(\cdot)$ approaches a standard Gaussian density ϕ as $\beta \to \infty$. To this end, with C and D denoting constants, and using the Taylor

expansion of $\log(1+x)$ about the point $x=0$ then

$$\begin{aligned}
\log\left(g_\beta(y)\right) &= C+\beta(a-1)\log\left(\mu+\frac{y}{\sqrt{-\beta h''(\mu)}}\right)-\beta b\left(\mu+\frac{y}{\sqrt{-\beta h''(\mu)}}\right)\\
&= D+\beta(a-1)\log\left(\left[1+\frac{y}{\mu\sqrt{-\beta h''(\mu)}}\right]\right)-\beta b\mu\left(\frac{y}{\mu\sqrt{-\beta h''(\mu)}}\right)\\
&= D+\beta(a-1)\left[\frac{y}{\mu\sqrt{-\beta h''(\mu)}}-\frac{y^2}{\mu^2(-\beta h''(\mu))^2}+O\left(\frac{1}{\beta^{3/2}}\right)\right]\\
&\quad -\beta b\left(\frac{y}{\sqrt{-\beta h''(\mu)}}\right)\\
&= D-\frac{y^2}{2}+O\left(\frac{1}{\beta^{3/2}}\right)\to\log(\phi(y))\quad\text{as}\quad\beta\to\infty.
\end{aligned}$$

Now, the aim is to compute $\mathrm{Var}_{g_\beta}\left(Y^2\right)$, and knowing the moments of $X\sim f^\beta$ which are easily attainable from the moment generator function of X, this can be done by using the fact that

$$Y=\sqrt{-\beta h''(\mu)}\,(X-\mu)\,.$$

Hence,

$$\begin{aligned}
\mathbb{E}_{g_\beta}\left[Y^2\right] &= \frac{\beta b^2}{a-1}\left[\mathbb{E}_\beta\left(X^2\right)-2\mu\mathbb{E}_\beta\left(X\right)+\mu^2\right]\\
&= \frac{\beta b^2}{a-1}\left[\frac{(\beta(a-1)+1)\,(\beta(a-1)+2)}{\beta^2b^2}\right.\\
&\qquad \left.-2\left(\frac{a-1}{b}\right)\left(\frac{\beta(a-1)+1}{\beta b}\right)+\frac{(a-1)^2}{b^2}\right]\\
&= \frac{\beta}{a-1}\left[\frac{3(a-1)}{\beta}+\frac{2}{\beta^2}-\left(\frac{2(a-1)}{\beta}\right)\right]\\
&= 1+\frac{2}{(a-1)\beta}\\
&= 1+O\left(\frac{1}{\beta}\right)
\end{aligned}$$

and letting $\alpha = \beta(a-1)+1$ for notational convenience then

$$\begin{aligned}
\mathbb{E}_{g_\beta}\left[Y^4\right] &= \frac{\beta^2 b^4}{(a-1)^2}\mathbb{E}_\beta\left((X-\mu)^2\right) \\
&= \frac{\beta^2 b^4}{(a-1)^2}\mathbb{E}_\beta\left[X^4 - 4X^3\mu + 6X^2\mu^2 - 4X\mu^3 + \mu^4\right] \\
&= \frac{\beta^2}{(a-1)^2}\left[\frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\beta^4} - 4\frac{\alpha(\alpha+1)(\alpha+2)(a-1)}{\beta^3}\right. \\
&\qquad \left. +6\frac{\alpha(\alpha+1)(a-1)^2}{\beta^2} - 4\frac{\alpha(a-1)^3}{\beta} + (a-1)^4\right] \\
&= 3 + \frac{26}{\beta(a-1)} + O\left(\frac{1}{\beta^2}\right)
\end{aligned}$$

and so

$$\mathrm{Var}_{g_\beta}\left(Y^2\right) = \mathbb{E}_{g_\beta}\left[Y^4\right] - \left(\mathbb{E}_{g_\beta}\left[Y^2\right]\right)^2 = 2 + O\left(\frac{1}{\beta}\right).$$

Hence, the assumption in equation (3.57) of Theorem 3.4.1 is satisfied with $\gamma = 1$.

t-distribution Example:

Suppose that $f(\cdot)$ is the density function of a t_ν-distribution (with ν degrees of freedom where $\nu > 4$). The density function is therefore given by

$$f(x) \propto \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The mode point for this distribution is $\mu = 0$, hence as in the statement of Theorem 3.4.1, $g_\beta(\cdot)$ is defined as

$$g_\beta(y) \propto f^\beta\left(\frac{y}{\sqrt{-\beta h''(0)}}\right)$$

which is achieved by transforming the original random variable, $X \sim f^\beta$, such that

$$X = \frac{Y}{\sqrt{-\beta h''(0)}}.$$

A routine calculation shows that

$$h''(0) = -\frac{\nu+1}{\nu}$$

and so

$$g_\beta(y) \propto \left(1 + \frac{y^2}{\beta(\nu+1)}\right)^{-\frac{\beta(\nu+1)}{2}}.$$

which can be recognised as the density function kernel of a scaled t-distribution. In fact Y can be written as $Y = \sigma' T$ where $T \sim t_{\nu'}$ where $v' = \beta(\nu+1) - 1$ and $(\sigma')^2 = \frac{\beta(\nu+1)}{\beta(\nu+1)-1} = \frac{\beta(\nu+1)}{\nu'}$.

Firstly, it will be shown that $g_\beta(\cdot)$ approaches a standard Gaussian density ϕ as $\beta \to \infty$. To this end, recalling that $\left(1 + \frac{x}{n}\right)^{-n} \to e^{-x}$ and $n \to \infty$,

$$\begin{aligned} g_\beta(y) &\propto \left(1 + \frac{y^2}{\beta(\nu+1)}\right)^{-\frac{\beta(\nu+1)}{2}} \\ &= \left(1 + \frac{y^2/2}{\beta(\nu+1)/2}\right)^{-\frac{\beta(\nu+1)}{2}} \\ &\to e^{-\frac{y^2}{2}} \quad \text{as} \quad \beta \to \infty. \end{aligned} \tag{3.70}$$

Now, the aim is to compute $\mathrm{Var}_{g_\beta}\left(Y^2\right)$. Recalling that the variance of a t-distribution on ν degrees of freedom is given by $\nu/(\nu-2)$, then

$$\begin{aligned} \mathbb{E}_{g_\beta}\left[Y^2\right] &= (\sigma')^2 \left[\frac{\nu'}{\nu'-2}\right] \\ &= \frac{\beta(\nu+1)}{\beta(\nu+1)-3} \\ &= 1 + \frac{3}{\beta(\nu+1)-3} \\ &= 1 + O\left(\frac{1}{\beta}\right). \end{aligned}$$

Recalling that the kurtosis of a random variable, Z, is defined as $\kappa_Z = \mathbb{E}\left(\frac{(Z-\mathbb{E}[Z])^4}{\mathrm{Var}(Z)^2}\right)$, and for a t-distribution on ν degrees of freedom the kurtosis is given by $3 + \frac{6}{\nu-4}$.

Hence,

$$\begin{aligned}
\mathbb{E}_{g_\beta}\left[Y^4\right] &= (\sigma')^4 \mathbb{E}\left[T_{\nu'}^4\right] \\
&= (\sigma')^4 [\mathrm{Var}\left(T_{\nu'}\right)]^2 \kappa_{T_{\nu'}} \\
&= \left(\frac{\beta(\nu+1)}{\nu'}\right)^2 \left[\frac{\nu'}{\nu'-2}\right]^2 \left[3+\frac{6}{\nu'-4}\right] \\
&= \frac{3\left[\beta(\nu+1)\right]^2}{\left(\beta(\nu+1)-3\right)^2} + \frac{6\left[\beta(\nu+1)\right]^2}{\left(\beta(\nu+1)-5\right)\left(\beta(\nu+1)-3\right)^2} \\
&= 3+\frac{2}{\beta(\nu+1)} + O\left(\frac{1}{\beta^2}\right)
\end{aligned}$$

and so

$$\mathrm{Var}_{g_\beta}\left(Y^2\right) = \mathbb{E}_{g_\beta}\left[Y^4\right] - \left(\mathbb{E}_{g_\beta}\left[Y^2\right]\right)^2 = 2 + O\left(\frac{1}{\beta}\right).$$

Hence, the assumption in equation (3.57) of Theorem 3.4.1 is satisfied with $\gamma = 1$. Consequently, in both the Gamma and t-distribution settings the result of Theorem 3.4.1 will hold for appropriately large β and with $\gamma = 1$. With k as defined in Theorem 3.4.1, the symmetry about the mode point in the t-distribution shows that $k = 3$; whereas the asymmetry of the Gamma distribution gives a $k = 5/2$ in that case.

The result in Theorem 3.4.1 does not imply that QuanTA isn't useful outside the Gaussian or super cold settings. The QuanTA approach will be practically useful in settings where the mode can be well approximated by a Gaussian and thus allow the shift move to approximately preserve the quantile. What Theorem 3.4.1 does show is that for a large class of distributions that exhibit appropriate smoothness, QuanTA is sensible, and is arguably the canonical approach to take at the super cold levels, since it enables acceleration of the mixing speed through the temperature schedule.

Chapter 4

Weight Preserved Tempering

4.1 Introduction

The PT/ST framework allows for any arbitrary density/mass function specification at the augmented (hot state) levels. In most applications the natural choice that maintains the location of the mode points is to raise the original target density to the power of an inverse temperature, β, i.e.

$$\pi_\beta(x) \propto f(x)^\beta. \tag{4.1}$$

Albeit the most easily implementable method to "spread out" the modes to allow for some overlap of modal mass, this can be a poor choice for the hotter state targets. The reason for this is that by powering up the target distribution, the relative weights of the multiple modes are not preserved in general. The weights of the modes can be significantly different in the hotter states than in the cold state target. Inevitably this leads to poor inter-modal mixing of the cold state chain between modes as the hotter states have weights, potentially drastically, inconsistent with those in the cold state target. Recall from Section 1.6 that the lack of regional weight preservation is one of the major features determining the torpid mixing of the PT algorithm in Woodard *et al.* [2009b].

The weight inconsistencies can prove highly misleading for the chains at the hotter temperatures. This "red-herring" effect is most obvious in a ST approach, where the chain explores the hotter levels and spends the vast majority of its time in regions that are relatively insignificant at the cold temperature. In fact, the chain often only "realises" that it has pursued a red-herring when it reaches the coldest levels; by which time it has wasted a large amount of time mixing through the temperature schedule without ever visiting the cold target state. An explicit

example of this effect will be given in this chapter in Section 4.3.1.

In Section 4.3.2 a new type of tempered target is proposed that (approximately) preserves modal weight in the canonical Gaussian setting and this leads to the proposal of a new prototype algorithm (the HAT algorithm) that attempts to overcome the weight inconsistency issues of the traditional PT scheme. This new scheme is highly computationally expensive and indeed has flaws for use in a real practical problems. However, it shows promise in the simulations in Section 4.4.1 and has already spawned further avenues for exploration and further work, see Chapter 6.

4.1.1 Heuristic Example

Consider the d-dimensional bimodal Gaussian target distribution with modes 1 and 2 with means, covariance matrices and weights given by μ_i, Σ_i, w_i for $i = 1, 2$ respectively. Hence the target is given by:

$$\pi(x) = Cf(x) = C\left(\frac{w_1}{(2\pi)^{\frac{d}{2}}|\Sigma_1|^{\frac{1}{2}}}\phi(x, \mu_1, \Sigma_1) + \frac{w_2}{(2\pi)^{\frac{d}{2}}|\Sigma_2|^{\frac{1}{2}}}\phi(x, \mu_2, \Sigma_2)\right), \quad (4.2)$$

where $\phi(x, \mu, \Sigma) = \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$.

Now assume that the hotter targets are generated through powering up the distribution as in equation (4.1). Furthermore assume that the modes 1 and 2 are well separated and thus have negligible mass between the modes. Then for β not too small and $\mathbf{x}$ appropriately close to mode i then the target can be approximated by

$$\begin{aligned}
\pi_\beta(x) &\propto \left(\frac{w_1}{(2\pi)^{\frac{d}{2}}|\Sigma_1|^{\frac{1}{2}}}\phi(x, \mu_1, \Sigma_1) + \frac{w_2}{(2\pi)^{\frac{d}{2}}|\Sigma_2|^{\frac{1}{2}}}\phi(x, \mu_2, \Sigma_2)\right)^\beta \\
&\approx \left(\frac{w_i}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}}\phi(x, \mu_i, \Sigma_i)\right)^\beta \\
&= \frac{w_i^\beta}{(2\pi)^{\frac{\beta d}{2}}|\Sigma_i|^{\frac{\beta}{2}}}\exp\left(-\frac{\beta}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right) \quad (4.3)
\end{aligned}$$

By integrating this (with the assumption that the modes are sufficiently spaced to

allow for the approximation to be valid) then

$$\begin{aligned}\int \pi_\beta(\mathbf{x})d\mathbf{x} &\approx C^\beta\Bigg[\int \frac{w_1^\beta}{(2\pi)^{\frac{\beta d}{2}}|\Sigma_1|^{\frac{\beta}{2}}}\exp\left(-\frac{\beta}{2}(\mathbf{x}-\mu_\mathbf{1})'\Sigma_1^{-1}(\mathbf{x}-\mu_\mathbf{1})\right)d\mathbf{x} \\ &\quad + \int \frac{w_2^\beta}{(2\pi)^{\frac{\beta d}{2}}|\Sigma_2|^{\frac{\beta}{2}}}\exp\left(-\frac{\beta}{2}(\mathbf{x}-\mu_\mathbf{2})'\Sigma_2^{-1}(\mathbf{x}-\mu_\mathbf{2})\right)d\mathbf{x}\Bigg] \\ &= C^\beta\left[\frac{w_1^\beta}{(2\pi)^{\frac{\beta d}{2}}|\Sigma_1|^{\frac{\beta}{2}}}\left(\frac{|\Sigma_1|}{\beta^d}\right)^{\frac{1}{2}} + \frac{w_2^\beta}{(2\pi)^{\frac{\beta d}{2}}|\Sigma_2|^{\frac{\beta}{2}}}\left(\frac{|\Sigma_2|}{\beta^d}\right)^{\frac{1}{2}}\right] \\ &= \frac{C^\beta}{((2\pi)^\beta\beta)^{\frac{d}{2}}}\left[w_1^\beta|\Sigma_1|^{\frac{1-\beta}{2}} + w_2^\beta|\Sigma_2|^{\frac{1-\beta}{2}}\right]. \end{aligned} \tag{4.4}$$

Denoting the weight of the i^{th} mode at the temperature level β by $W_{(i,\beta)}$ then the following approximation can be made

$$\begin{aligned} W_{(i,\beta)} &\approx \frac{\frac{C^\beta}{((2\pi)^\beta\beta)^{\frac{d}{2}}}\left[w_i^\beta|\Sigma_i|^{\frac{1-\beta}{2}}\right]}{\frac{C^\beta}{((2\pi)^\beta\beta)^{\frac{d}{2}}}\left[w_1^\beta|\Sigma_1|^{\frac{1-\beta}{2}} + w_2^\beta|\Sigma_2|^{\frac{1-\beta}{2}}\right]} \\ &= \frac{w_i^\beta|\Sigma_i|^{\frac{1-\beta}{2}}}{w_1^\beta|\Sigma_1|^{\frac{1-\beta}{2}} + w_2^\beta|\Sigma_2|^{\frac{1-\beta}{2}}} \\ &\propto w_i^\beta|\Sigma_i|^{\frac{1-\beta}{2}}, \end{aligned} \tag{4.5}$$

where the final line is taking proportionality with respect to the i^{th} mode's mass. Consequently, for β not too small the target can be approximated by

$$\pi_\beta(x) \propto W_{(1,\beta)}\,|\Sigma_1|^{-\frac{1}{2}}\,\phi(x,\mu_1,\Sigma_1/\beta) + W_{(2,\beta)}\,|\Sigma_2|^{-\frac{1}{2}}\,\phi(x,\mu_2,\Sigma_2/\beta). \tag{4.6}$$

Now consider the case when the target is a one-dimensional bimodal Gaussian of the form given in equation (4.2) with parameters: $\mu_1 = -40, \mu_2 = 40, \Sigma_1 = 1, \Sigma_2 = 9, w_1 = 0.9$ and $w_2 = 0.1$. Hence

$$\pi(x) \propto \frac{0.9}{(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x+40)^2\right) + \frac{0.1}{(2\pi)^{\frac{1}{2}}\times 3}\exp\left(-\frac{1}{2\times 9}(x-40)^2\right). \tag{4.7}$$

Figure 4.1 illustrates the inconsistency of the modal weights at the different temperature levels when power tempering is used. In fact, as the $\beta \to 0$ then it is clear that the mode centred at 40 begins to dominate the share of the weight even though this mode in the cold state is only attributable for a weight of 0.1.

Visually from Figure 4.1 it is clear that there is inconsistency of the weights

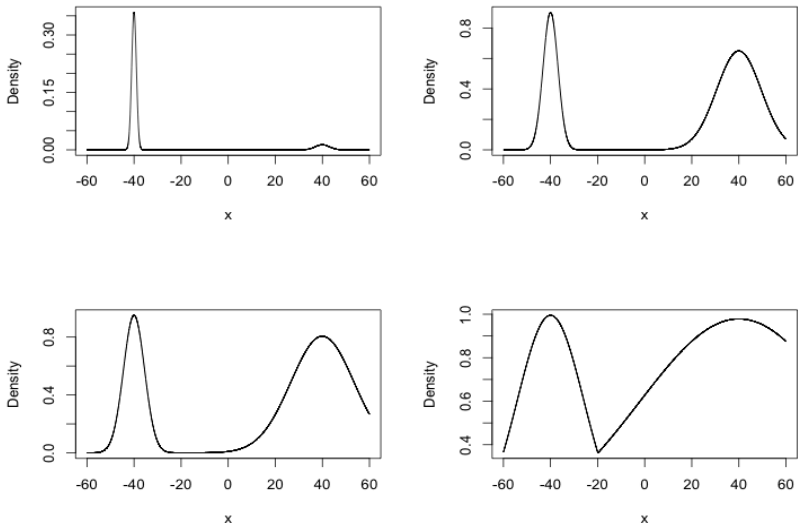

Figure 4.1: Plots of the tempered target densities generated by powering the target, $\pi(x)$, given in equation (4.7) by powers $\beta = \{1, 0.1, 0.05, 0.005\}$. It is clear from the plot that the relative weights in the modes in the first plot at the cold temperature with $\beta = 1$ are not preserved and in fact the cold state lower weight mode increasingly dominates the share of the weight as $\beta \to 0$

at the hotter and colder temperatures. Equation (4.5) can be used to approximate the weights at the temperature level β for this example

$$W_{(1,\beta)} \approx \frac{0.9^{\beta} \times 1^{1-\beta}}{0.9^{\beta} \times 1^{1-\beta} + 0.1^{\beta} \times 3^{1-\beta}} \tag{4.8}$$

and

$$W_{(2,\beta)} \approx \frac{0.1^{\beta} \times 3^{1-\beta}}{0.9^{\beta} \times 1^{1-\beta} + 0.1^{\beta} \times 3^{1-\beta}}. \tag{4.9}$$

β	1	0.1	0.05	0.005
$W_{(1,\beta)}$	0.90	0.32	0.28	0.25
$W_{(2,\beta)}$	0.1	0.68	0.72	0.75

Table 4.1: Approximated weights associated with modes 1 and 2 from the target distribution given in equation (4.7) and computed using the formulas in equations (4.8) and (4.9).

Table 4.1 gives the approximate weights of the modes at the levels $\beta =$

$\{1, 0.1, 0.05\}$. It confirms the observations regarding the relative weights in the pair of modes from Figure 4.1. It is apparent from the values in Table 4.1 that mode 2 becomes the dominant mode of the pair as the temperature increases. Even when moving from the cold state to the neighboring state at $\beta = 0.1$, mode 1 reduces its share of the global mass from 90% to just 32%.

Consequently, when the parallel tempering algorithm is used with power based tempering then the chains at the hotter states are targeting distributions that can have modal weights significantly different to those in the cold state. As will be discussed in Section 4.1.2, this leads to reduced swap acceptance probabilities for useful swap moves; a direct consequence of the hotter states suggesting the "wrong" modal weights when proposing a swap location for the chain.

4.1.2 The Effect on the Swap Move Acceptance Probabilities

The effects of using power-based targets from equation (4.1) on the swap acceptance rates is best understood through the analysis of the swap acceptance ratio when considering the bimodal Gaussian target example given in equation (4.2). So consider two particles $\mathbf{x}$ and $\mathbf{y}$ at the inverse temperature levels β and $\beta^{'}$ respectively and suppose that $\mathbf{x}$ is currently in mode 1 and $\mathbf{y}$ is currently in mode 2. Furthermore, assume that the two temperatures levels are still cold enough that the modes are still well separated. Now suppose that a swap move has been proposed between these two levels.

The Metropolis-Hastings swap move ratio is (approximately) given by

$$\begin{aligned} \alpha &= \frac{\pi(\mathbf{x})^{\beta^{'}}\pi(\mathbf{y})^{\beta}}{\pi(\mathbf{y})^{\beta^{'}}\pi(\mathbf{x})^{\beta}} \\ &\approx \left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)^{\frac{\beta-\beta^{'}}{2}} \times \frac{\exp\left(-\frac{\beta^{'}-\beta}{2}(\mathbf{x}-\mu_{\mathbf{1}})'\Sigma_1^{-1}(\mathbf{x}-\mu_{\mathbf{1}})\right)}{\exp\left(-\frac{\beta^{'}-\beta}{2}(\mathbf{y}-\mu_{\mathbf{2}})'\Sigma_2^{-1}(\mathbf{y}-\mu_{\mathbf{2}})\right)}. \end{aligned} \tag{4.10}$$

Equation (4.10) shows that there is a clear dependence on the respective covariance structures of the modes.

An intuitive analysis is given when the two key terms in equation (4.10) are considered separately. The two distinct terms in this ratio are:

$$A_1 := \left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)^{\frac{\beta-\beta^{'}}{2}} \quad \text{and} \quad A_2 := \frac{\exp\left(-\frac{\beta^{'}-\beta}{2}(\mathbf{x}-\mu_{\mathbf{1}})'\Sigma_1^{-1}(\mathbf{x}-\mu_{\mathbf{1}})\right)}{\exp\left(-\frac{\beta^{'}-\beta}{2}(\mathbf{y}-\mu_{\mathbf{2}})'\Sigma_2^{-1}(\mathbf{y}-\mu_{\mathbf{2}})\right)}. \tag{4.11}$$

As has already been motivated in Chapter 1 in Section 1.5 and Chapter 2, the

ideal would be to have the swap ratio, $\alpha \approx A_1 A_2$, as close to 1 as possible for an arbitrarily large temperature spacing and thus optimise the mixing rate of the algorithm through the temperature schedule.

Chapter 2 introduced QuanTA, which utilises a reparametrisation move that is motivated by preservation of the quantile of a Gaussian mode upon tempering. This seeks to nullify the effects of shrinkage/expansion of the mode through the tempering schedule. In fact QuanTA controls of the A_2 term to be equal to 1 in this Gaussian setting. However, this does not solve the whole problem and it is apparent that the A_1 term also has a significant effect on the performance of the algorithm.

It is immediately apparent from the form of A_1 that the acceptance ratio, α, degrades exponentially quickly in the inverse temperature spacing with a rate proportional to the log ratio of the covariance structures of the respective modes.

Recall the setting given above in equation (4.7). In this case, suppose that the two particles to be swapped are indeed in separate modes, so a "useful" swap has been proposed, then for any choice of β and $\beta^{'}$ in the range that keeps the approximations valid

$$A_1 = \left(\frac{1}{3}\right)^{\beta - \beta^{'}} \tag{4.12}$$

and so when considering the spacings analysed in Table 4.1 then for the levels at $\beta = 1$ and $\beta^{'} = 0.1$ then $A_1 = 0.37$. So even if QuanTA is used to preserve the quantile within a mode, making $A_2 \approx 1$, then the move will have an acceptance rate far from 1.

Monitoring acceptance rates doesn't tell the whole story though. The above only explains the degradation of the swap rate for moves in the case when the two particles are in different modes. If the locations of the two chains are within the same mode then irrespective of the change in weight of the mode between levels, the acceptance probability will remain high, with A_1 being equal to 1 in such cases. This is particularly problematic since the chains can become trapped in regions that dominate the mass at the hot temperature levels; essentially only visiting these regions and thus giving reasonable acceptance rates. This wouldn't be diagnosed by checking the swap acceptance rates since few if any "useful" temperature swap moves are being performed.

Therefore, temperature swap acceptance rates are therefore generally not a reliable diagnostic, something noted in Woodard *et al.* [2009b]. This is certainly something that a practitioner should be aware of, particularly when running the algorithms in high dimensions. Concrete examples of when the acceptance rates appear acceptable but in fact there has been an absolute failure to mix are given in

both Sections 4.3.1 and 4.4.1.

4.2 The Ideal Tempering Targets

The above motivates searching for a version of the tempered targets that preserves the regional weight but still provides the inter-modal bridging mass. Note that this section only focuses on the Gaussian mixture setting as this will be a useful approximation to many more general settings.

The modal weight preservation would mean that the chains at the hotter states would be in the correct regions with consistent regularity to the cold state target. This should enable even more ambitious proposals throughout the temperature schedule since the algorithm no longer has to overcome the weight inconsistencies.

Consequently, although obviously impractical, the ideal solution would be to adjust the dispersion in the mode in the same way that power based tempering of a uni-modal target would; but now whilst maintaining the modal weights.

The best way to describe this would be to refer back to the Gaussian mixture example in the motivating example with density given in equation (4.2). In the power based tempering case then the target distribution at inverse temperature level β is given by

$$\pi_\beta(\mathbf{x}) \propto \left(\frac{w_1}{(2\pi)^{\frac{d}{2}}|\Sigma_1|^{\frac{1}{2}}}\phi(\mathbf{x}, \mu_1, \Sigma_1) + \frac{w_2}{(2\pi)^{\frac{d}{2}}|\Sigma_2|^{\frac{1}{2}}}\phi(\mathbf{x}, \mu_2, \Sigma_2) \right)^\beta . \tag{4.13}$$

However, as seen in Section 4.1.1, this does not generally preserve the relative weights of the modes at the hotter temperatures.

Modal weight would be preserved if, instead, at temperature level β, the target distribution was given by

$$\pi_\beta(\mathbf{x}) \propto \frac{w_1}{(2\pi)^{\frac{d}{2}}|\Sigma_1|^{\frac{1}{2}}}\phi\left(\mathbf{x}, \mu_1, \frac{\Sigma_1}{\beta}\right) + \frac{w_2}{(2\pi)^{\frac{d}{2}}|\Sigma_2|^{\frac{1}{2}}}\phi\left(\mathbf{x}, \mu_2, \frac{\Sigma_2}{\beta}\right) . \tag{4.14}$$

Using this as the hotter state target would indeed preserve the regional weights whilst also providing the ultimate goal of modal dispersion. Considering the swap acceptance ratio between two consecutive chains, located in two separate modes at levels β and β', as was done in Section 4.1.2 in equation (4.10), but now

for this idealised target then

$$\begin{aligned}\alpha &= \frac{\pi_{\beta'}(\mathbf{x})\pi_{\beta}(\mathbf{y})}{\pi_{\beta'}(\mathbf{y})\pi_{\beta}(\mathbf{x})} \\ &\approx \frac{\exp\left(-\frac{\beta'-\beta}{2}(\mathbf{x}-\mu_{\mathbf{1}})'\Sigma_1^{-1}(\mathbf{x}-\mu_{\mathbf{1}})\right)}{\exp\left(-\frac{\beta'-\beta}{2}(\mathbf{y}-\mu_{\mathbf{2}})'\Sigma_2^{-1}(\mathbf{y}-\mu_{\mathbf{2}})\right)} = A_2. \end{aligned} \tag{4.15}$$

The A_1 term from equation (4.15) has now disappeared and consequently has no effect on the swap acceptance. Thus, if used in conjunction with the QuanTA approach, making $A_2 \approx 1$, then there would be super fast mixing through the (cooler) parts of the temperature schedule.

Figure 4.2 shows the comparison between the target distributions used when using the power based setup vs the idealised setup for the example density given in equation (4.7). The densities have been normalised for ease of comparison and clearly illustrate how the mode centred at 40 progressively becomes the dominant weight mode as the power based tempering scheme is used.

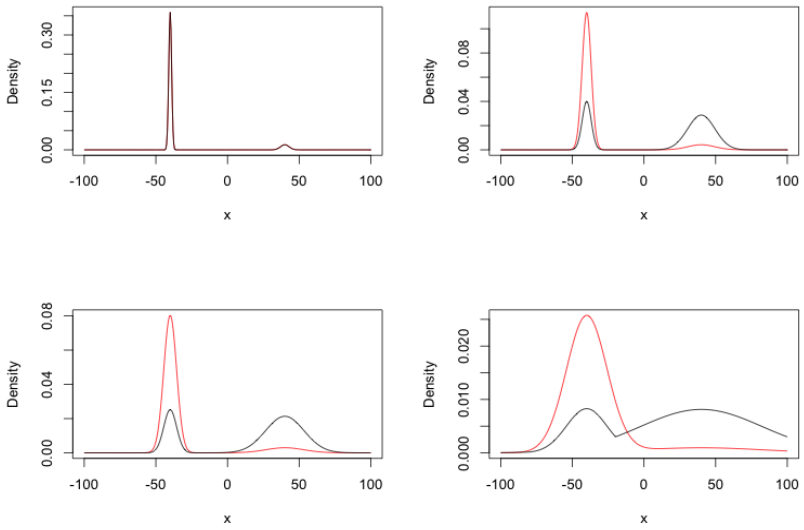

Figure 4.2: Plots of the normalised tempered target densities generated by both powering the target (black) and the ideal modal spread (red). $\pi(x)$ as in equation (4.7) and inverse temperature levels $\beta = \{1, 0.1, 0.05, 0.005\}$.

Hence, in the practically unrealistic setting where the target is a Gaussian mixture with known parameters, including the weights, then the idealised target

is taken to be the mixture with a tempered variance in each component. Thus, spreading out the modes but preserving the regional weights.

To formalise this, for a general Gaussian mixture target given by

$$\pi(x) \propto \sum_{j=1}^{J} w_j \phi(x, \mu_j, \Sigma_j) \tag{4.16}$$

then the corresponding idealised target distribution is defined as:

Definition 4.2.1 (Idealised Target Distribution)**.** For a Gaussian mixture target distribution $\pi(\cdot)$, as in equation (4.16), the idealised tempered target at inverse temperature level β is defined as

$$\pi_\beta^I(x) \propto \sum_{j=1}^{J} w_j \phi\left(x, \mu_j, \frac{\Sigma_j}{\beta}\right). \tag{4.17}$$

Using these idealised targets in the PT scheme can give substantially better performance than when using the standard power based targets. This is illustrated in the examples in Sections 4.3.1 and 4.4.1. In practice the ideal target will need to be approximated; this will be reviewed in Section 4.3.2.

From herein, when the term "Ideal Algorithm" is used it refers to the implementation of the standard PT algorithm but now using the idealised targets from equations (4.17).

4.3 The Impact of High Dimensionality

So far this issue with the weightings has only been illustrated in a one-dimensional target scenario. Recall that in the Gaussian mixture setting with well separated modes then the acceptance rate for a swap move given by equation (4.10) depends directly on two terms A_1 and A_2. In a Gaussian mixture setting the A_2 term can be made to be approximately 1 when QuanTA is applied to the problem. This leaves the acceptance probability equal to

$$A_1 = \left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)^{\frac{\beta-\beta'}{2}}.$$

This quantity decays/grows exponentially with the temperature spacing between consecutive levels. There is also significant decay with the dimensionality increase. To get a basic understanding of the impact of dimensionality then suppose the

covariance matrices Σ_1 and Σ_2 have iid structure with marginal variances σ_1^2 and σ_2^2 respectively. Then in this case

$$A_1 = \left(\frac{\sigma_1}{\sigma_2}\right)^{d(\beta-\beta')} \tag{4.18}$$

and so the acceptance ratio of a swap move grows/decays exponentially quickly in dimension, d. This is a key feature that leads to the torpid mixing of parallel tempering in high dimensions that is described in detail in Woodard *et al.* [2009b]; specifically the exponential decay of the persistence quantity found in Woodard *et al.* [2009b].

4.3.1 A Warning Example of Naively Using Power Tempering in High Dimensions

In the following there will be a basic example comparing the performances of the **simulated** tempering in the two cases when the naive power based targets are used and then when the toy "Idealised" targets in Section 4.2 are used instead. A simplistic higher dimensional bimodal setting will illustrate that when standard power based targets are used at the hotter states then:

1. Acceptance rates cannot be relied upon to diagnose poor inter-modal mixing of the chain;

2. Even in a 10-dimensional problem the standard power based tempering scheme leads to critically bad performance of the simulated tempering algorithm.

Use of the ST algorithm as opposed to the PT algorithm is due to ease of analysis and understanding. There are no multiple chain interactions in the ST algorithm. A key observation will be to understand the bottlenecks that hinder the algorithm, and these are only obvious in this simulated tempering setup.

The example considered is the ten dimensional target distribution given by the bimodal Gaussian mixture

$$\pi(x) = w_1\phi_{(\mu_1,\Sigma_1)}(x) + w_2\phi_{(\mu_2,\Sigma_2)}(x) \tag{4.19}$$

where $w_1 = 0.2$, $w_2 = 0.8$, $\mu_1 = (-10, -10, \ldots, -10)$, $\mu_2 = (10, 10, \ldots, 10)$, $\Sigma_1 = 9\mathbf{I}_{10}$ and $\Sigma_2 = \mathbf{I}_{10}$. Hence, at the hotter temperature levels when power based tempering is used then mode 1, which only accounts for 20% of the mass at the cold level, becomes the dominant mode containing almost all the mass.

For both runs the same geometric temperature schedule was used:

$$\{1, 0.32, 0.32^2, \ldots, 0.32^6\}.$$

In fact, Theorem 5.1.1 from Chapter 5 suggests this is an optimal setup for the Idealised run of the algorithm since the swap move acceptance rates are around 0.22; close to the suggested 0.234 optimal value. However, as will be seen, this will not be optimal for the run using the power based targets despite the algorithm having a not unreasonable 0.17 swap acceptance rate between the coldest and next coldest levels. This is a little on the low side and this is due to the fact that the targets at all levels have been normalised after power based tempering. This is unrealistic in practice and instead an adaptive normalisation approach such as that in Atchadé and Liu [2004] would be needed. Obviously, this is why the parallel tempering approach is more accessible and practical but the interaction of particles in a swap move can hide the the key observations in this example. Both algorithms were run from an initial location which is the modal point of the mode 1 region of mass.

Another key part of the setups of the algorithm was the toy setup for the algorithms' within temperature proposals. In order to ensure that the within modal mixing isn't influencing the mixing then a local modal independence sampler was used for the within moves. This essentially means that once a mode has been found the mixing is "infinitely" fast. Using the naive modal assignment that the location $\mathbf{x}$ is in mode 1 if $\bar{\mathbf{x}} < 0$ and in mode 2 otherwise, then the within move proposal distribution for a move at inverse temperature level β is given by

$$q_\beta(\mathbf{x}, \mathbf{y}) = \phi_{\left(\mu_1, \frac{\Sigma_1}{\beta}\right)}(\mathbf{y}) \mathbb{1}_{\bar{\mathbf{x}}<0} + \phi_{\left(\mu_2, \frac{\Sigma_2}{\beta}\right)}(\mathbf{y}) \mathbb{1}_{\bar{\mathbf{x}}\geq 0}, \tag{4.20}$$

where $\phi_{\mu,\Sigma}(.)$ is the density function of a Gaussian random variable with mean μ and variance matrix Σ.

Figure 4.3 plots a functional of the inverse temperature at each iteration of the algorithm runs. The functional is

$$h(\beta_t)\text{sgn}\left(\bar{\mathbf{x}}_t\right) := \frac{\log\left(\frac{\beta_t}{\beta_m}\right)}{\log\left(\frac{1}{\beta_m}\right)}\text{sgn}\left(\bar{\mathbf{x}}_t\right) \tag{4.21}$$

where sgn(.) is the usual sign function and β_m is the minimum of the inverse temperatures. The key here is that the levels are now being illustrated on the log scale with sign indication of the modal location. Without the log scale for this geometri-

cally decaying schedule it would be very unclear what is happening at the hottest temperatures.

Figure 4.3 clearly illustrates that the hot state modal weight inconsistency leads the chain down a red-herring trajectory with the suggestion that all the mass is in modal region 1; resulting in the chain **never** reaching the other mode in the finite run of the algorithm. This is highlighted further in Figure 4.4, which shows the estimated kernel density estimates in the two cases for the marginal distribution of the first component.

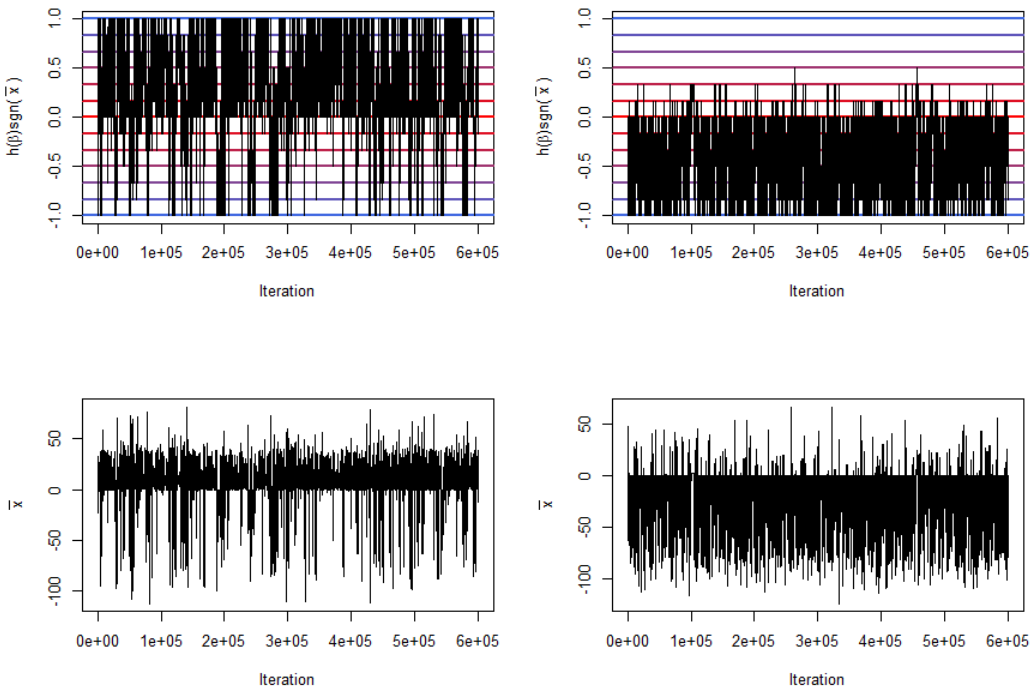

Figure 4.3: Top: Trace plots of the functional of the simulated tempering chains given in equation (4.21). On the left is the version using the idealised targets, which mixes well through the temperature schedule and finds both modal regions. On the right is the version using the standard power-based targets, which fails to ever find one of the modes. Bottom: Trace plots showing the associated trace plot of $\bar{x}$ in the two cases.

0.234 Rule Failure:

In this case there was evidence that the power-based target runs were sub optimally tuned since the swap acceptance rate between the coldest and second

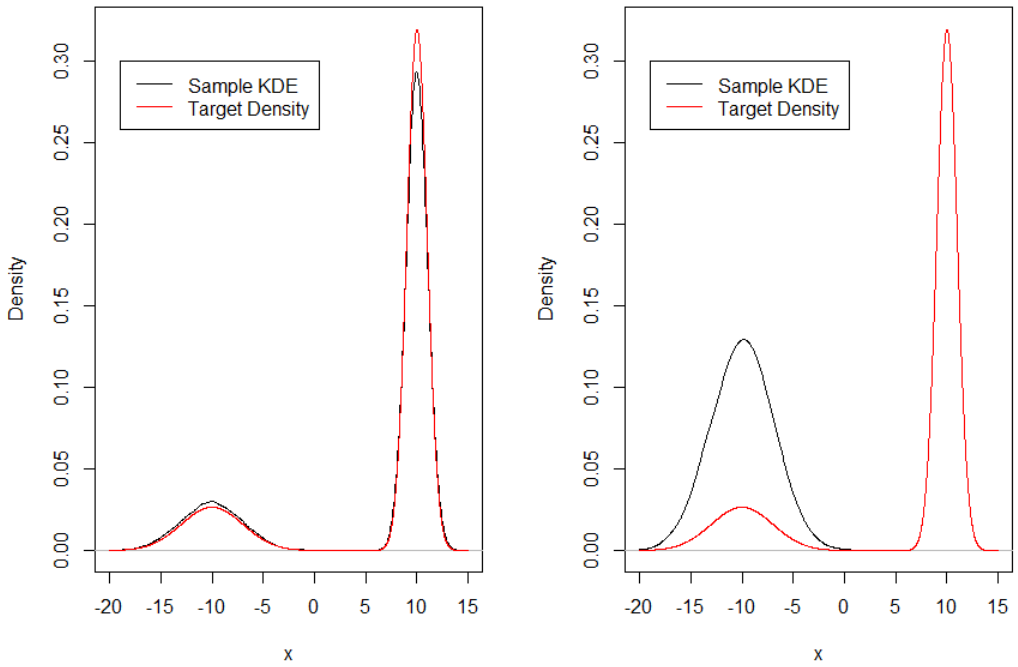

Figure 4.4: Figure 4.3 showed the relative performances of the simulated tempering runs using idealised and power-based versions of the target respectively. To further highlight the gains of using the idealised targets, here are the respective kernel density estimates from the generated samples from the two runs.

coldest levels was on the low side. This is only noticeable given the (unrealistic) normalisation of the temperature marginals and indeed would be un-noticeable if the PT algorithm was employed.

Even so, there was certainly no indication from the swap ratios between any of the other levels that there was poor performance in the power-based target setup; all these swaps were close to the suggested optimal 0.234 rate. A closer examination of the 0.234 rule statement in Atchadé *et al.* [2011] will explain why this "optimal setup" still allowed poor mixing.

Atchadé *et al.* [2011] showed that when the target distribution has a global iid form, $\pi(\mathbf{x}) \propto \prod_{i=1}^{d} f(x_i)$, then the optimal temperature spacings in a traditional power-based algorithm are geometric when

$$I(\beta) = \mathrm{Var}_{f^\beta}[\log(f(x))] \propto \beta^{-2}. \tag{4.22}$$

This is the case when the target distribution is the uni-modal multivariate Gaussian target with iid marginals.

The temperature spacings in Section 4.3.1 were designed to be optimal for the idealised targets. This schedule was itself geometric and appeared to be the correct scaling even for the power-based targets since it gave approximately 0.234 swap acceptance rates throughout the schedule. This gives the first bit of insight into why the setup failed to be optimal. For a Gaussian mixture under power-based tempering, the optimal schedule would not generally be geometric since the result in equation (4.22) wouldn't hold. But in the simulation run the geometric schedule gave close to the optimal acceptance rate, showing that essentially the run has been tuned to be optimal in a single Gaussian mode rather than across the multiple modes.

Indeed, the trace plots in Figure 4.3 show that the chain is effectively trapped in mode 1, which although it only has 20% of the mass in the cold state, is completely dominant at the hotter states. Consequently, in finite runs starting in mode 1, aiming for a 0.234 swap rate is essentially tuning the algorithm to mix in a uni-modal Gaussian since the chain is never reaching the other mode.

This ultimately highlights the issue of using the metric of Expected Squared Jumping Distance, Atchadé *et al.* [2011], or the limiting diffusion speed, Roberts and Rosenthal [2014], in conjunction with the assumption of "infinitely fast" within temperature mixing to obtain results suggesting optimality of the tempering schedule.

The key problem is that the suggested optimal schedules are derived under the (practically) unrealistic assumption that the chains can mix (globally) infinitely fast in each of the temperature levels. This isn't an unrealistic assumption when the target distribution is unimodal or in the case when there is multimodality but where all the modes have a symmetric form and equal weights, see Woodard *et al.* [2009a].

However, in the cases when the modes have different scalings, as was the case in the example in Section 4.3.1, maximising the $ESJD_\beta$ in the temperature space proves to be a misleading metric to optimise; especially if infinitely fast within temperature mixing is assumed. Particularly in the context of PT algorithm, in the setting of Atchadé *et al.* [2011], there is no consideration of what constitutes a "useful swap move". A non-rigorous definition of what it means to be a useful swap move is given later in Definintion 5.2.1. Essentially it constitutes a swap move in a PT algorithm between chains that are located in different modes.

Without this notion, tuning the spacings to have a 0.234 acceptance rate for

the temperature level swaps in a finite run of the algorithm could simply be tuning the algorithm to only work well in a single mode. This gives the chain the "optimal" ability to mix through the temperature schedule but without necessarily exchanging worthwhile mixing information.

As a preview, Section 5 derives optimal scaling results for a regionally weight preserving PT algorithm and discusses the reliability of using the swap move acceptance rates as a diagnostic for algorithmic performance. Section 5.2 builds on this by suggesting potential improvements to the temperature swap rate diagnostics.

A heuristic calculation gives further intuition to the issues with the power-based optimal setup in Atchadé *et al.* [2011] and Roberts and Rosenthal [2014]. Suppose that the d-dimensional state space, $\mathcal{X}_d$, is made up of a disjoint union of two regions $A_{(1,d)}$ and $A_{(2,d)}$ and that the target distribution has the regionally conditionally independent identically distributed form

$$\pi(\mathbf{x}) = \sum_{k=1}^{2} \left[\prod_{i=1}^{d} f_k(x_i) \right] \mathbb{1}_{[\mathbf{x} \in A_{(k,d)}]}. \tag{4.23}$$

Furthermore assume that the regions are given by hyper-rectangles such that $A_{(k,d)} = A_k^1 \otimes \ldots \otimes A_k^d = [a_k, b_k] \otimes \ldots \otimes [a_k, b_k]$. In the case of power based tempering, then at inverse temperature level β, the weight in the k^{th} region w_k^β is given by

$$w_k^\beta \propto \int_{A_{(k,d)}} \left[\prod_{i=1}^{d} f_k^\beta(x_i) \right] d\mathbf{x} = \left(\int_{a_k}^{b_k} f_k^\beta(z) dz \right)^d. \tag{4.24}$$

Now define the ratio of weights in the two regions at the inverse temperature level β to be $r(\beta) = w_1^\beta / w_2^\beta$. The ratio of these ratios at consecutive temperature levels β and $\beta' = \beta + \epsilon$ will be considered in the high dimensional setting when ϵ is necessarily small. Using a Taylor expansion to the first order term

$$\log(r(\beta')) - \log(r(\beta)) = \epsilon \frac{\partial}{\partial \beta} \left[\log(r(\beta)) \right] + O(\epsilon^2), \tag{4.25}$$

where it can be shown that

$$\frac{\partial}{\partial \beta} \left[\log\left(r(\beta)\right) \right] = d \left[\mathbb{E}_{f_1^\beta} \left[\log\left(f_1\right) \right] - \mathbb{E}_{f_2^\beta} \left[\log\left(f_2\right) \right] \right]. \tag{4.26}$$

The relevance of this is that in the optimal scaling setting of Atchadé *et al.* [2011] it is fundamental that the spacings of consecutive inverse temperature levels have a dimensionality scaling $\epsilon = \ell / d^{1/2}$. This dimensionality scaling is essential to achieving a non-degenerate asymptotic swap acceptance rate.

In contrast, the above heuristic in equations (4.24), (4.25) and (4.26) show that for $\epsilon \propto 1/d^{1/2}$ then $r(\beta')$ becomes exponentially inconsistent with $r(\beta)$ as dimensionality increases. This shows that the setting considered in Atchadé *et al.* [2011] is too simplistic to explain the scaling issues prevalent in high-dimensional complex settings.

4.3.2 Approximating the Ideal

Using the idealised targets given in Section 4.2 would be typically impossible since weights, locations and scales of the modes are unknown. The idealised targets effectively assumes knowledge of the relative modal weights apriori. This is unrealistic, but this section introduces a prototype method that is designed to approximate these idealised targets and maintain some of the benefits of the idealised targets in low to mid-dimensional settings.

Essentially one would like to remove the weight inconsistencies through the temperature schedule. At temperature level β then the approximate weight of the i^{th} mode is given by $W_{(i,\beta)}$ in equation (4.5), where it is approximated as

$$W_{(i,\beta)} \propto w_i^{\beta} |\Sigma_i|^{\frac{1-\beta}{2}}. \tag{4.27}$$

Now assume that the target distribution is (or at least can be well approximated by) a d-dimensional Gaussian mixture and so the target distribution is given by

$$\pi(x) \propto f(x) = \sum_{i=1}^{N} w_i \phi_i(x) \tag{4.28}$$

where w_i is the (unnormalised) weight of the i^{th} mode and ϕ_i is the Gaussian density of the i^{th} mode with mean μ_i and covariance matrix Σ_i. Assume that the modes are well spaced. Tempering this with the traditional power based method can have major issues as described in Section 4.1.1.

Using the ideal tempering targets of Section 4.2 then consider, at inverse temperature level β, targeting the a "weight adjusted" target

$$\pi_\beta(x) \propto f(x)^\beta \alpha_\beta(x) \tag{4.29}$$

where function, $\alpha_\beta(x)$, is designed to "preserve" the modal weight throughout the temperature schedule. Thus $\alpha_\beta(\cdot)$ is acting as a multiplicative correction factor for the traditional power based target given by $f(\cdot)^\beta$.

Let x be a point in the i^{th} mode which has local mean μ_i and covariance

structure Σ_i. Appealing to the weight inconsistency issue highlighted above in equation (4.27) and the assumption that the modes are well spaced then to preserve the weight the ideal form of $\alpha_\beta(x)$ is

$$\alpha_\beta(x) = w_i^{1-\beta}|\Sigma_i|^{\frac{\beta-1}{2}}. \tag{4.30}$$

This would then mean that the approximate weight, $W_{(i,\beta)}^{\text{Adj}}$, of the i^{th} mode at level β for the adjusted density π_β^{Adj} is given by

$$W_{(i,\beta)}^{\text{Adj}} \propto W_{(i,\beta)} \times \alpha_\beta(x) = w_i \tag{4.31}$$

as required.

In general the w_i are unknown and in fact are values that are being estimated and so a surrogate must be used instead.

Assuming that the mode is approximately a unimodal Gaussian then by considering the normalisation constant of the unimodal mode we can get values proportional to the true weights. Hence consider an unnormalised d-dimensional unimodal Gaussian density $f(y)$ with normalisation constant C, mean μ and covariance structure Σ. The normalisation constant, C, can be calculated as

$$C = f(y) \times (2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}} \times \exp\left\{\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}. \tag{4.32}$$

Recalling that if the distribution is indeed Gaussian then at any location y

$$\nabla \log\big(f(y)\big) = -\Sigma^{-1}(y-\mu) \tag{4.33}$$

and

$$\nabla\nabla^T \log f(y) = -\Sigma^{-1}. \tag{4.34}$$

Hence using equations (4.33) and (4.34) then

$$\begin{aligned} C &= f(y) \times (2\pi)^{\frac{d}{2}}|(-\nabla\nabla^T \log f(y))^{-1}|^{\frac{1}{2}} \\ &\quad \times \exp\left\{\frac{1}{2}\nabla \log\big(f(y)\big)^T \left[\nabla\nabla^T \log f(y)\right]^{-1} \nabla \log\big(f(y)\big)\right\}. \end{aligned} \tag{4.35}$$

Returning to the multimodal setting and now using the weight surrogate given in equation (4.35), then especially in the Gaussian mixture setting, given in

equation (4.28), an appropriate choice of the adjustment function $\hat{\alpha}_\beta$ is given by

$$\begin{aligned}
\hat{\alpha}_\beta(x) &= \left[f(x) \times (2\pi)^{\frac{d}{2}} \, |(-\nabla\nabla^T \log f(x))^{-1}|^{\frac{1}{2}} \right. \\
&\quad \left. \times \exp\left\{ \frac{1}{2} \nabla \log\big(f(x)\big)^T \left[\nabla\nabla^T \log f(x) \right]^{-1} \nabla \log\big(f(x)\big) \right\} \right]^{1-\beta} \\
&\quad \times |(-\nabla\nabla^T \log f(x))^{-1}|^{\frac{\beta-1}{2}} \\
&= (2\pi)^{\frac{d(1-\beta)}{2}} f(x)^{(1-\beta)} \\
&\quad \times \exp\left\{ \frac{1-\beta}{2} \nabla \log\big(f(x)\big)^T \left[\nabla\nabla^T \log f(x) \right]^{-1} \nabla \log\big(f(x)\big) \right\} \quad (4.36)
\end{aligned}$$

With the adjustment in equation (4.36) then the naive adjusted target can be defined as

Definition 4.3.1 (Numerically Adjusted Target)**.** For a target distribution, $\pi(\cdot)$ in C^2, the corresponding adjusted target is defined as

$$\begin{aligned}
\pi_\beta^{\text{Adj}}(x) &\propto f(x)^\beta \hat{\alpha}_\beta(x) \\
&= (2\pi)^{\frac{d(1-\beta)}{2}} f(x) \exp\left\{ \frac{1-\beta}{2} \nabla \log\big(f(x)\big)^T \left[\nabla\nabla^T \log f(x) \right]^{-1} \nabla \log\big(f(x)\big) \right\} \\
&\propto f(x) \exp\left\{ \frac{1-\beta}{2} \nabla \log\big(f(x)\big)^T \left[\nabla\nabla^T \log f(x) \right]^{-1} \nabla \log\big(f(x)\big) \right\}. \quad (4.37)
\end{aligned}$$

In its raw form this is quite an elegant formula and one can immediately see from the form given in equation (4.37) that the adjusted target is effectively the original cold state target multiplied by a Gaussian inspired term that contains all the tempering features for "spreading out the mass". Suppose that f was indeed a Gaussian unimodal density then the adjusted target reassuringly gives $\pi_\beta^{\text{Adj}}(x) \propto f(x)^\beta$.

4.3.3 Problem Points for the Adjusted Target

Assuming well-spaced modes that are approximately unimodal Gaussian within each mode then this adjusted density appropriately attempts to replicate the "ideal targets" of Section 4.2. However, in the low density areas between the modes the raw form of the adjustment given in equation (4.37) behaves very badly. This is due to the use of the hessian of the logged target (see equation (4.34)) at each location being used as an approximation to the covariance structure of the local mode. Note that in the canonical Gaussian mixture case then this will be a very good approxi-

mation when the location is within a mode, however into the tails and in particular in the zone in between the modes then the hessian can suggest a covariance matrix that is not positive definite.

In fact at the points of inflection of the logged target density between the modes then the value of $\left[\nabla\nabla^T \log f(x)\right]^{-1}$ is very unstable. In the one-dimensional case the adjusted target given in equation (4.37) explodes at these points resulting in an ill-defined target distribution. Thus using the naive targets defined in equation (4.37) can have serious stability issues.

4.3.4 A Robust Adjusted Target

In the naive approach given in equation (4.37), the local gradient information is used to associate a particle with a mode and then provides a multiplicative adjustment factor for the weight re-adjustment.

Furthermore, in a Gaussian mixture setting, note that at the mode points of the modes when the gradient terms are 0 then equation (4.37) shows that the local mode heights are being preserved. Hence, the adjusted target can essentially be seen as a local rescaling of the powered target distribution. In fact, suppose that the target is a uni-modal Gaussian distribution with mean μ and variance matrix Σ. Then, in this case, it can be seen that the naive adjusted target in equation (4.37) is equivalent, up to proportionality, to

$$\pi_\beta^{\text{Adj}}(x) \propto f(x)^\beta f(\mu)^{1-\beta} \tag{4.38}$$

which has the interpretation that the mode height is being preserved throughout tempering by repeatedly rescaling according to how the mode point's height changes as the target is tempered.

Extending this to a Gaussian mixture setting, note that:

1. At inverse temperature level, β, and locally to a mode the adjusted target, $\pi_\beta^{\text{Adj}}(\cdot)$, is (up to proportionality) the usual tempered Gaussian derived from power-based tempering.

2. Consider the form of $\pi_\beta^{\text{Adj}}(\cdot)$ given in the final line of equation (4.37). One can see that at the any given mode point μ, when the gradient terms are 0, then up to a global proportionality constant $\pi_\beta^{\text{Adj}}(\mu) \propto f(\mu)$, with no β dependency on the RHS.

Hence, the adjusted target can essentially be seen as a local rescaling of the powered target distribution in such settings. Moreover, if one only rescales regions using a

localised formula similar to equation (4.38) for $\beta \in [0, 1)$ then there will be no explosion points that would render the target distribution improper.

Thus the aim is to associate each point x with a mode point, μ_x to which equation (4.38) will then be applied. Essentially providing the desired, localised rescaling.

In the Gaussian uni-modal setting the mode point can be found exactly through the gradient information at the current location, x by using just one step of the Newton optimisation algorithm such that

$$\mu = x - \left[\nabla\nabla^T \log f(x)\right]^{-1} \nabla \log\big(f(x)\big). \tag{4.39}$$

In a Gaussian mixture setting this approach is suitable for the majority of points in the state space to associate a location to a local mode point. Although further work is needed to establish a robust approach outside this setting, for the prototype algorithm in this thesis it is sufficient to define

$$\mu_x = x - \left[\nabla\nabla^T \log f(x)\right]^{-1} \nabla \log\big(f(x)\big). \tag{4.40}$$

Then, assuming that each point, x, in the state space has an associated mode point, μ_x, the robust adjusted target at tempering level β is defined as

Definition 4.3.2 (Robust Adjusted Target)**.** Consider a target distribution $\pi(x) \propto f(x)$ in C^2 then the robust adjusted target is defined as

$$\pi_\beta^H(x) \propto f(x)^\beta f(\mu_x)^{1-\beta}. \tag{4.41}$$

In the form given in equation (4.41), the height of the adjusted target at all levels is bounded by the global maximum value of the target distribution and hence there is no longer a concern that there will be explosion points as was the case in the numerically adjusted target given in Definition 4.3.1.

In the regions where the hessian is not strictly negative definite the value of μ_x given by the single step of the Newton scheme in equation (4.40) will not necessarily be a position that will have a higher density value. In fact, for many of the values in this region then the value of μ_x will be out into the tails of the target and so rescaling to the height of the target at μ_x will effectively keep zero density (rather than explosion) at such points.

Consider the example in equation (4.7) given earlier for a bimodal univariate mixture of Gaussians. Figure 4.5 illustrates the behaviour of the robust adjusted target relative to the idealised target; showing that until the hottest temperatures

levels are reached the numerical version is a good approximation to the idealised target.

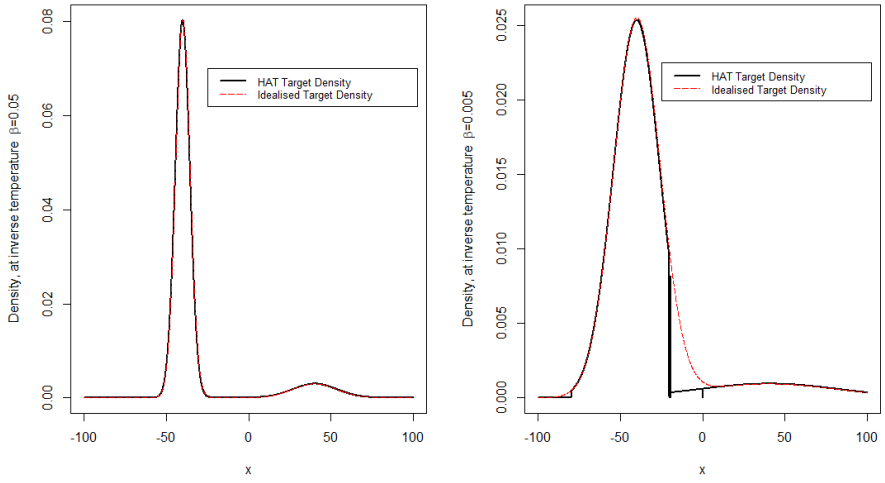

Figure 4.5: Plot of the robust vs ideal tempered targets for the bimodal Gaussian example in equation (4.7) at inverse temperatures $\beta = 0.05$ and $\beta = 0.005$ respectively. Note the almost identical behaviour at the colder temperature; but the step change in the hotter temperature, which is an open issue with the adjusted targets.

At this point one should check that this robust adjusted target is indeed a well-defined probability distribution at inverse temperature level β.

Proposition 4.3.1. *Assume that $\pi(x) = Cf(x)$ is continuous and bounded on $\mathbb{R}^d$ and that*

$$\int_{\mathcal{X}} f^{\beta}(x)dx < \infty.$$

Then $\pi_{\beta}^{H}(x)$ is a well defined probability density.

Proof. By integrability of $f^{\beta}(\cdot)$, there exists a $C \in \mathbb{R}$ such that $\int_{\mathcal{X}} f(x)dx = \frac{1}{C} < \infty$. By boundedness of $\pi(\cdot)$ there exists an $M \in \mathbb{R}$ such that$M = \sup_x\{f(x)\} < \infty$. Then, at inverse temperature level β

$$\begin{aligned}\int_{\mathcal{X}} \pi_{\beta}^{H}(x)dx &= \int_{\mathcal{X}} Cf(x)^{\beta} f(\mu_x)^{1-\beta}dx \\ &\leq \int_{\mathcal{X}} Cf(x)^{\beta} M^{1-\beta}dx \\ &= CM^{1-\beta}\int_{\mathcal{X}} f(x)^{\beta}dx < \infty. \end{aligned} \tag{4.42}$$

So $\pi_\beta^H(x) \geq 0$ is finitely integrable. □

4.3.5 Improvements to the Adjusted Target

Section 4.3.4 introduces the new, robust adjusted tempered target distributions that in the canonical Gaussian mixture setting (approximately) preserve modal mass. Recall that the computation of $\pi_\beta^H(x)$ at each point requires the association of x to a local modal point μ_x that will, hopefully, be the mode point of the local mode. In the Gaussian setting this can be found using a single step of the Newton optimisation scheme. However, this is still an issue for points in between modes where the Hessian is not strictly negative definite. From herein this region where the hessian of the logged target is not strictly negative definite will be referred to as the **zone of uncertainty**.

Gaussian mixtures are the focus of the examples in this chapter. However, two avenues that are furtherwork for generalisation/improvement of the numerically adjusted target are:

1. A Quasi-Newton scheme could be used, so that points even in the zone of uncertainty can be allocated to a suitable mode point.

2. A more general, but fundamentally deterministic, optimisation scheme could be used e.g. gradient ascent.

4.4 The HAT (Hessian Adjusted Tempering) Algorithm

Sections 4.3.2, 4.3.5 and 4.3.4 gave suggestions for a new type of tempered target where, especially, in the Gaussian mixture setting in the colder temperatures the regional weights are at least approximately maintained. Hence the adjusted target at inverse temperature level β is specified to be

$$\pi_\beta^H(x) \propto f(x)^\beta f(\mu_x)^{1-\beta} \tag{4.43}$$

where μ_x is derived deterministically from the location x using either

1. A single step of a Newton optimisation scheme (requiring hessian calculation);

2. Use the first option but additionally use the methods discussed in Section 4.3.4 to add mass to the zone of uncertainty (but this should only be done at the hotter temperatures, for details see Section 4.6);

3. Use another local optimisation scheme such as Gradient Ascent, which would be necessary when the target isn't C^2.

The proposal is that one uses the adjusted target suggested in the form given in equation (4.43) along with added robustness suggested in options 2 and 3 succeeding equation (4.43). Consequently the actual algorithm is just that of the vanilla PT algorithm given in Section 1.4.2 but instead using the new adjusted targets. Even so the algorithm is given for completeness:

HAT (Hessian Adjusted Tempering) algorithm :

- Choose a sequence of inverse temperature values $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$. This should be done with guidance from the optimal spacing strategy suggested in Theorem 5.1.1 and the Corollaries 5.2.1 and 5.2.2 found in Chapter 5.
- Choose initial values of the chains for each temperature level, $x_0^0, x_1^0, \ldots, x_n^0$.
- Choose the proposal mechanisms for the within temperature level moves at each level, $q_{\beta_j}(x_i^j, x_{(i+1)}^j)$ for $j = 0, 1, \ldots, n$. See Section 4.6 regarding within temperature moves at the hottest levels.
- Choose the number, m, of within temperature proposals the chains will perform before attempting a swap type move and choose the total number, s, of swap moves that will be attempted.
- Then **iterate s times**:
    1. Perform m within temperature moves for each of the $(n+1)$ chains according to the chosen proposal mechanism at each level; maintaining invariance with respect to the adjusted target given in equation (4.43). For the hotter states, especially in the higher dimensional settings, then more care should be taken to ensure fast mixing, see Section 4.6.
    2. Uniformly randomly select a pair of adjacent inverse temperatures, β_j and β_{j+1} say, for which a swap move is proposed, and where the values of the respective chains are (currently) x_j and x_{j+1}.
    3. Compute the acceptance ratio for the proposed swap and accept the swap with probability equal to

$$\min\left(1, \frac{\pi^H_{\beta_{j+1}}(x_j)\pi^H_{\beta_j}(x_{j+1})}{\pi^H_{\beta_j}(x_j)\pi^H_{\beta_{j+1}}(x_{j+1})}\right).$$

- End, and discard a suitably chosen burn in period for the chain.

4.4.1 Examples of Implementation of the HAT Algorithm

The effectiveness of the "Ideal" weight preserving targets has already been illustrated in Section 4.2. Through a series of examples it will be shown that, even with the HAT approximation, in the canonical Gaussian setting there is comparable performance to the ideal algorithm.

4.4.2 One-dimensional Gaussian mixture example:

Consider a bi-modal Gaussian mixture target with target density given by:

$$\pi(x) \propto \sum_{k=1}^{2} w_k \phi_{(\mu_k, \sigma_k^2)}(x) \tag{4.44}$$

where $\phi_{(\mu,\sigma^2)}(.)$ is the density function of a univariate Gaussian with mean μ and variance σ^2. The modal weights are given by $w_1 = 0.8$ and $w_2 = 0.2$, the means are given by $\mu_1 = -40$ and $\mu_2 = 40$ and finally the standard deviations are given by $\sigma_1 = 0.1$ and $\sigma_2 = 5$. So there is a large disparity between the modal variances. Figure 4.6 illustrates the target distribution π and clearly shows the variance and height disparity between the modes.

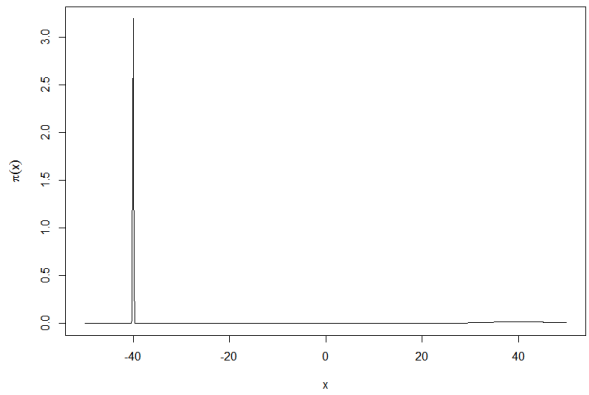

Figure 4.6: The target density plotted for the example bi-modal target given in equation (4.44). Note that the second mode located at 40 is very disperse yet is the dominant mode at the hotter temperatures under vanilla power tempering.

The performance of the new HAT algorithm will be compared with that of the PT algorithm. However it is not obvious how to construct a fair setup since the "optimal temperature schedules" will differ between the two algorithms. Both have a suggested optimal 0.234 rule for the acceptance rates of the spacings (see Section 5 for details of the optimal scaling of the HAT algorithm).

To illustrate the gains of the HAT algorithm in this case the same temperature schedule will be used and this will be chosen under optimality for the HAT algorithm. This will highlight that the optimal spacings for HAT algorithm in this example are too ambitious for the PT algorithm to work well.

The inverse temperature schedule used is geometrically spaced with common ratio 0.05. Hence, the inverse temperature schedule is $\{1, 0.05, 0.05^2\}$. Verified over 10 repeated runs of the HAT algorithm this schedule is (approximately) optimal for the HAT algorithm, according to Theorem 5.1.1 from Chapter 5. Furthermore, at each level there are three within temperature moves before a temperature swap proposal between a uniformly selected pair of consecutive temperatures. The run is finished when there have been 20,000 swap moves proposed. All chains were started at the position 40 to really highlight the lack of robustness of the PT algorithm to a feasible start point.

Figure 4.7 shows three trace plots, of the cold state chain, for runs of the PT algorithm under the described setup. The target weight in the modal region centred on -40 is 0.8; so it is clear that the performance of the PT algorithm is inconsistent, making modal weight estimates highly variable.

Figure 4.8 shows the trace plot of three runs of the cold state chain for the HAT algorithm. It is immediately obvious that the inter-modal mixing is far more regular than for the PT approach. The acceptance rates of the swap moves between the coldest and consecutively next coldest level for each of these three runs are respectively $\{0.27, 0.29, 0.28\}$; for comparison those of the PT runs in Figure 4.7 were $\{0.05, 0.12, 0.12\}$. In this case, which is simplistic due to the single dimension, the poor mixing in the target state in the PT algorithm would be picked up by the low swap move acceptance rate and hence in an optimal setup further intermediate temperatures would be needed. This illustrates the ability of the HAT algorithm to take larger steps through the temperature schedule than the PT algorithm is capable of. In the below example, in five dimensions there are illustrations of the swap acceptance rates not diagnosing the poor mixing.

Figure 4.9 shows the running modal weight approximation of w_1 after the k^{th} iteration of the cold state chains once a burn-in period of 10,000 iterations has been removed for the ten examples of the PT and HAT runs respectively. The weight

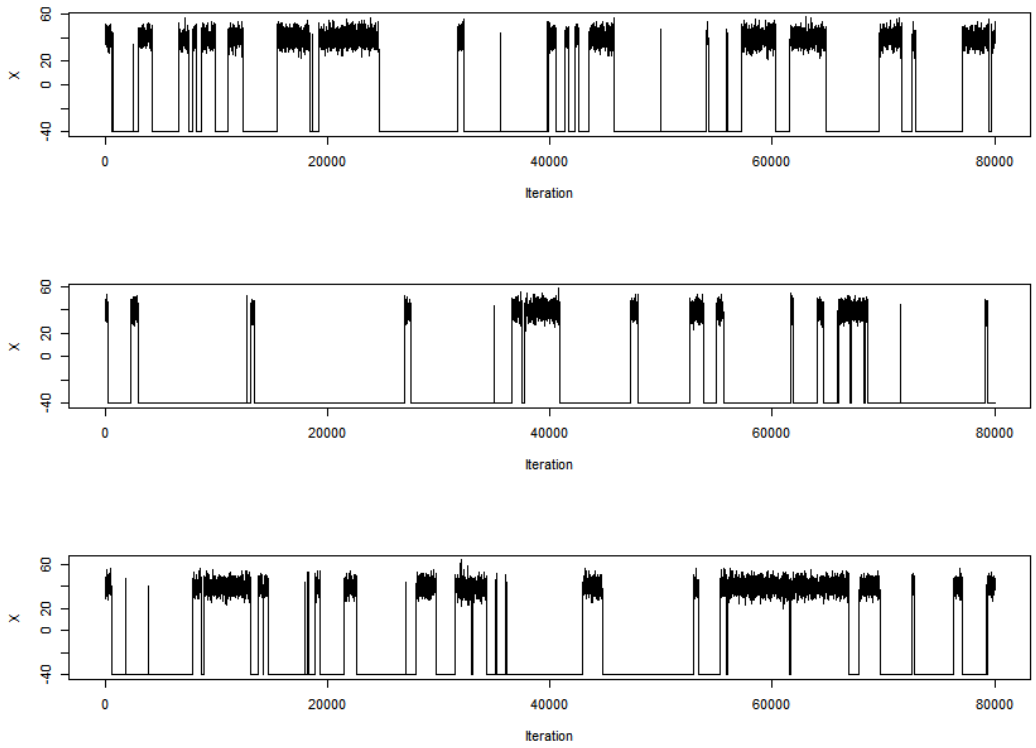

Figure 4.7: Three trace plots of the mixing of the coldest level chain in three separate runs of the PT algorithm targeting the distribution given in equation (4.44). The setup of the PT algorithm was the same in each case and the key observation is the infrequency of inter modal jumps that would subsequently result in more variable estimates of modal weights.

approximation after iteration k is given by

$$\hat{w}_1^k = \frac{1}{k - 10000} \sum_{i=10001}^{k} \mathbb{1}_{(X_i<0)} \tag{4.45}$$

where X_i is the location of the chain at the coldest temperature level after the i^{th} iteration. Observe the jagged and volatile convergence of the running estimate of $\hat{w}_1^k$ as it converges to the true value 0.8 for the PT algorithm.

To see how the performance of the HAT algorithm compares with that of the idealised algorithm in this Gaussian mixture setting then 10 runs of the Ideal algorithm were performed. All runs had the same setup with regards to the temperature

Figure 4.8: Three trace plots of the mixing of the coldest level chain in three separate runs of the HAT algorithm targeting the distribution given in equation (4.44). The setup of the HAT algorithm was the same in each case and the key observation is the relatively high frequency of inter modal jumps which one would hope would give a fast rate of convergence of an estimator of the modal weights.

schedule and within level performance. The runs gave comparable performance to the HAT version. An example comparing a run from each of the three types is given in Figure 4.10. It is hard to differentiate between the trace plots for the HAT and idealised runs.

4.4.3 Five-dimensional example:

Consider the target distribution again given by the bimodal Gaussian mixture. Consider a bi-modal Gaussian mixture target with target density given by:

$$\pi(x) \propto \sum_{k=1}^{2} w_k \phi_{(\mu_k, \Sigma_k)}(x) \tag{4.46}$$

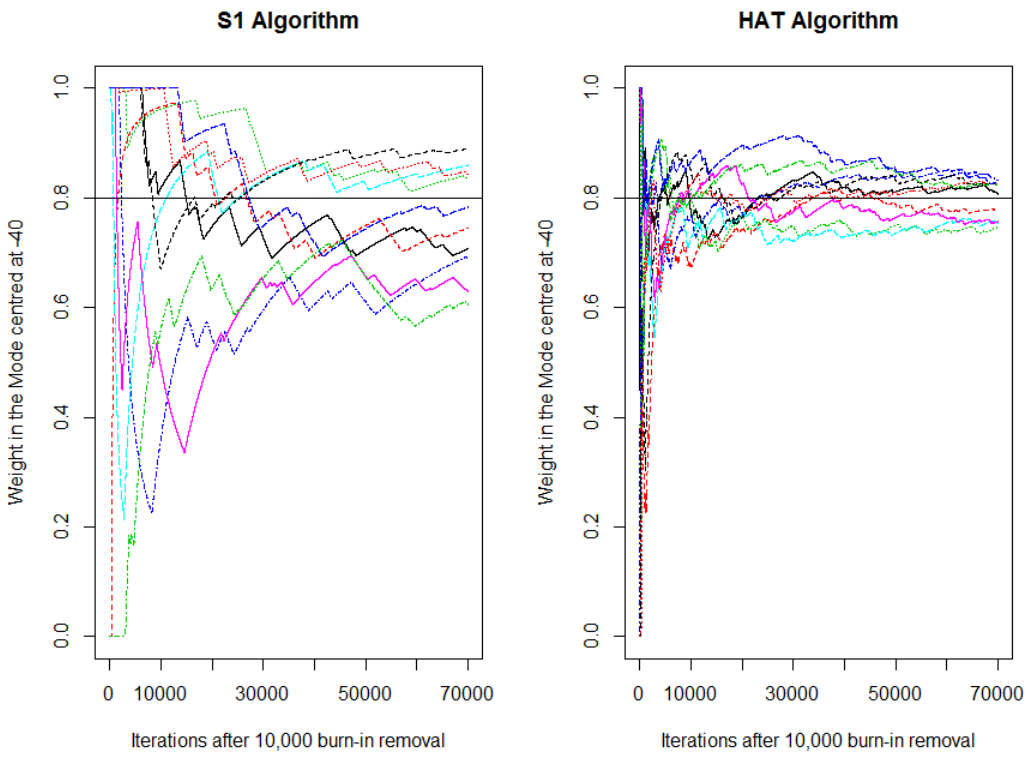

Figure 4.9: Running estimate of the weight in the mode centred on -40, from equation (4.45), for 10 runs of the PT (here denoted S1) and HAT algorithms respectively. In both cases a burn-in of 10,000 iterations was removed. Observe the increased variability of the weight estimates for the PT runs compared to the HAT runs.

where $\phi_{(\mu,\Sigma)}(.)$ is the density function of a 5 dimensional Gaussian with mean μ and covariance matrix Σ. The weights of the modes are even with $w_1 = w_2 = 0.5$, $\mu_1 = (-15, \ldots, -15)$, $\mu_2 = (15, \ldots, 15)$, $\Sigma_1 = I_5$ and $\Sigma_2 = 3^2 \times I_5$. This example will be used to analyse the following:

1. Similarly to the one-dimensional example, the performance of the HAT algorithm will be compared to the PT algorithm under a setup of optimality for the HAT algorithm.

2. There will be a comparison of performance when the problem is made only slightly harder with more ambitious temperature spacings that will illustrate the added robustness of the HAT scheme over the PT algorithm in this setting.

3. A basic examination of the hot state mixing which will be used to motivate the discussion in Section 4.6.

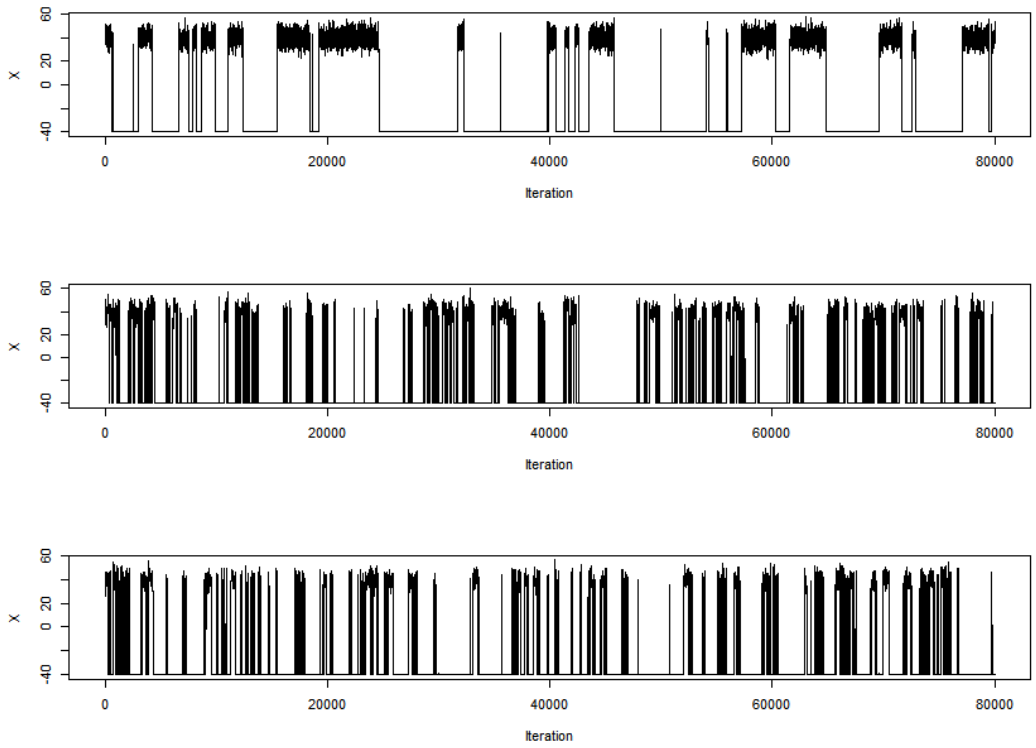

Figure 4.10: Three trace plots of the cold state chain targeting the distribution in equation (4.44) using the PT, HAT and Ideal algorithms respectively. Note the visually comparable performance between the HAT and Ideal runs.

Firstly, appealing to the results in Theorem 5.1.1 and Corollary 5.2.1 from Chapter 5, then the optimal inverse temperature schedule for the HAT setup on seven levels is geometrically spaced and is given by $\{1, 0.35, 0.35^2, \ldots, 0.35^6\}$. Running this over 10 runs gives stable empirical estimates of the swap ratios at around 0.26 for the HAT algorithm, which only defer from this value in the hottest levels when the weight preserving approximation becomes less valid. The algorithm is again setup with 3 within temperature level moves and then one proposal of a temperature swap. Both modes have identical weights and in all simulations a fixed start location of $(15, \ldots, 15)$ was used; obviously this has a biasing effect from the start but it is useful to show that the PT algorithm either fails or takes a very long time to burn-in from a very reasonable start point.

Figure 4.11 shows three runs of the PT algorithm on this setup. As was apparent in the one-dimensional case the mixing is extremely poor with very few

transitions between modes. There is a clear problem with the algorithm struggling to escape the burn-in period. Clearly any attempt to estimate regional weights from these finite runs would be heavily biased.

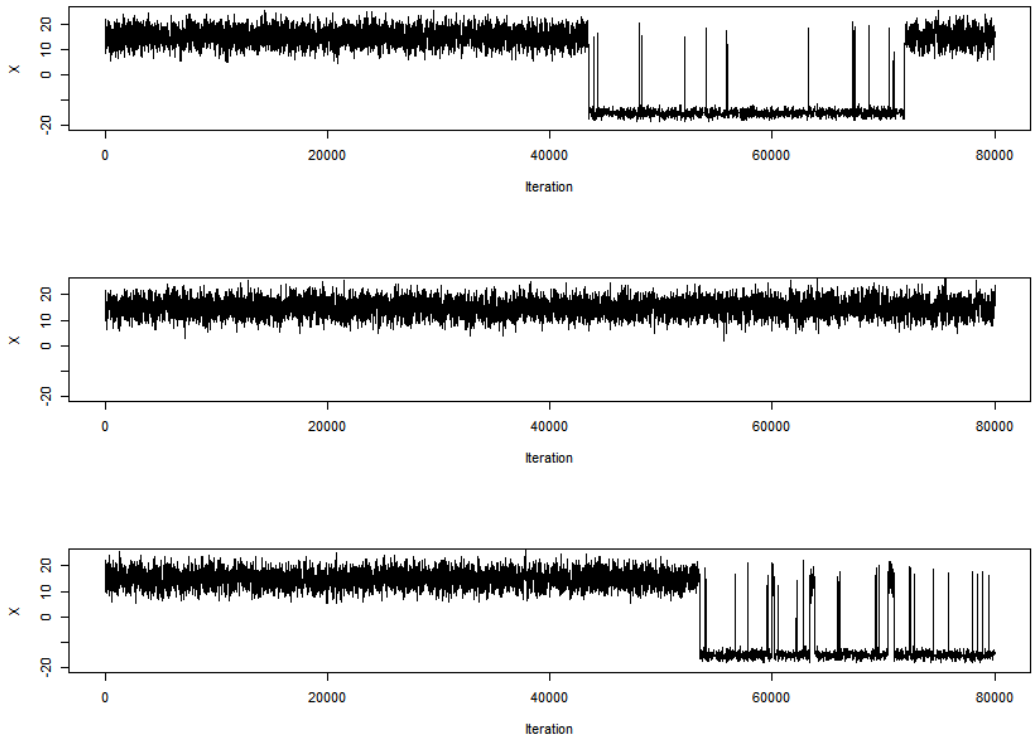

Figure 4.11: Three trace plots of the first component mixing of the coldest level chain in three separate runs of the PT algorithm targeting the distribution given in equation (4.46). There is a clear infrequency of inter-modal jumps. The swap move acceptance rates between the coldest state and the next level in the three cases are $\{0.18, 0.27, 0.21\}$.

On the other-hand Figure 4.12 shows three runs of the HAT algorithm implementation on its optimised setup. Relative to the mixing in the PT algorithm the frequency of inter-modal moves at the cold state is much higher and unlike the volatile performance of the PT algorithm the runs on this finite number of iterations, at least by eye, appear to be consistently good. The regularity of swap moves means that once a suitable (but relatively small) burn-in period has been removed then the runs of the HAT algorithm provide far more stable and lower variance estimates of the modal weights.

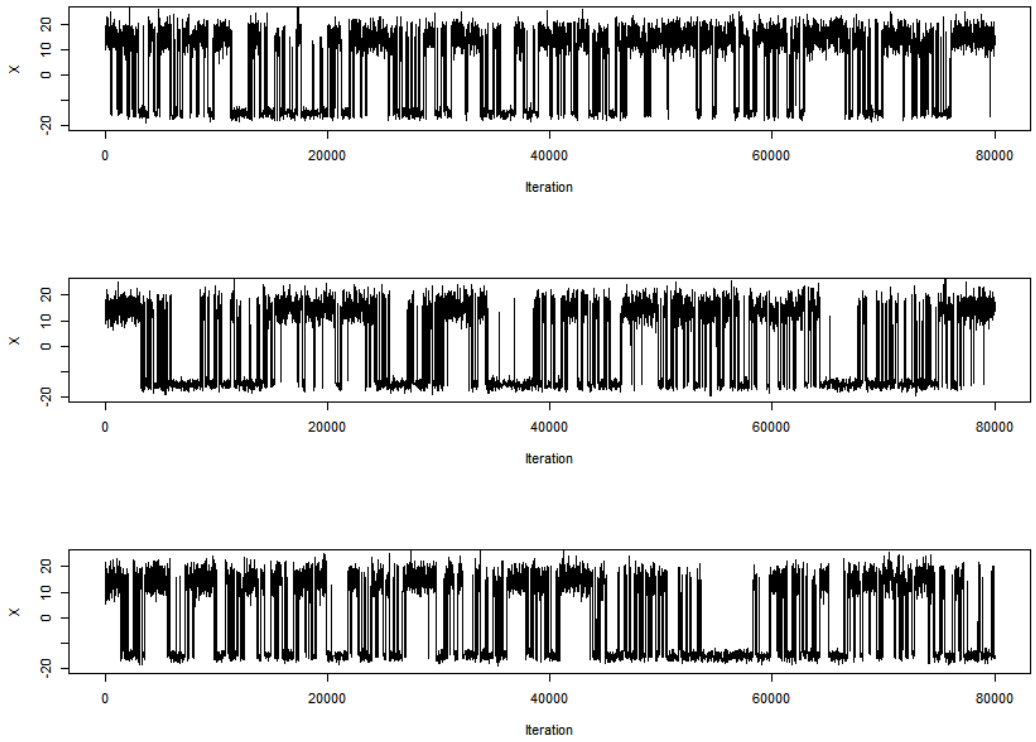

Figure 4.12: Three trace plots of the first component mixing of the coldest level chain in three separate runs of the HAT algorithm targeting the distribution given in equation (4.46). Note the high frequency of inter modal jumps.The swap move acceptance rates between the coldest state and the next level in the three cases respectively are $\{0.27, 0.28, 0.27\}$.

Figure 4.13 shows one example of a run of each of the algorithms in this setup, i.e. the PT, HAT and Ideal. Although just one example is given, it is hoped that this shows that there is comparability in the trace plots of the performance of the HAT algorithm to the performance of the Ideal algorithm.

Next, there is a brief look at the hot state mixing of the algorithms for the runs given in Figure 4.14. The mixing for the hot state of the PT algorithm is very fast and the chain is able to move around the state space quickly. However, this is not the case for the HAT algorithm which suffers from some aspect of modality still. The problem with the mixing here is discussed more deeply in Section 4.6 and is one of the key areas of further work proposed from this thesis.

Finally, a slight increase in the ambitiousness of the spacing is made. The

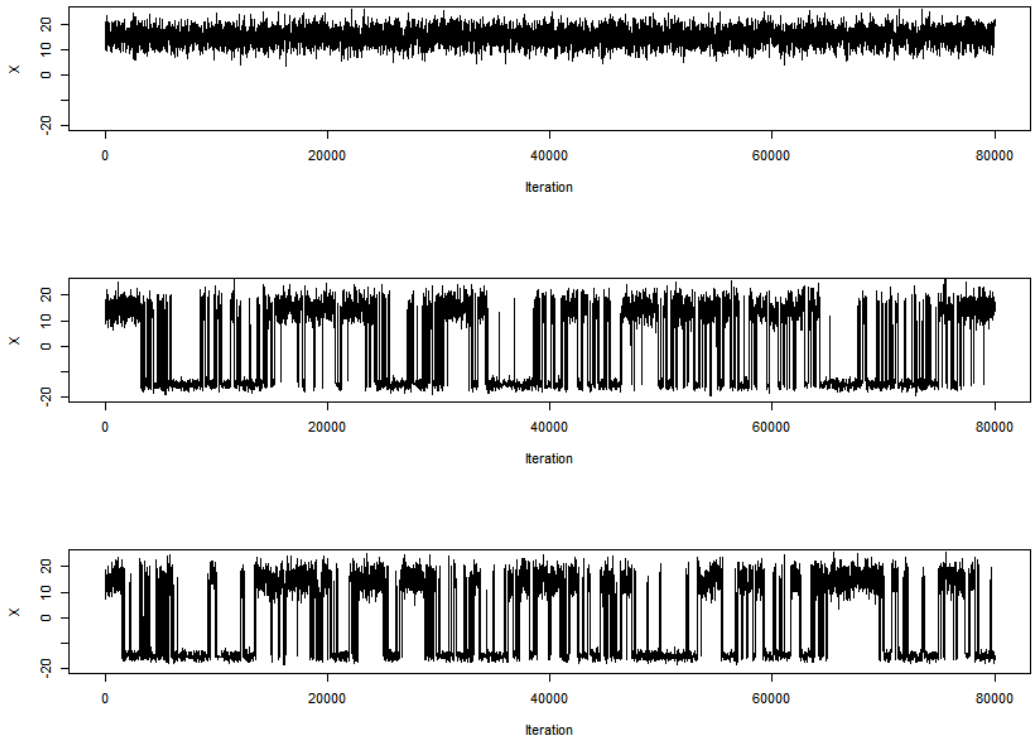

Figure 4.13: Three trace plots of the cold state chain targeting the distribution in equation (4.44) using the PT, HAT and Ideal algorithms respectively.

common ratio of the geometric spacing decreases from 0.35 to 0.25. Figure 4.15 shows one representative run from each. The PT algorithm in these finite runs never finds the other mode and, although slightly on the low side, the acceptance rate for swap moves throughout the temperature schedule is approximately 0.16 and would not be suggestive of any major mixing issue to the practitioner. The acceptance rates too for the HAT algorithm in this case are also only 0.16, again a little on the low side from that suggested for optimality, but this is not totally prohibitive for the algorithm which still manages to make, albeit less but still, regular inter-modal jumps.

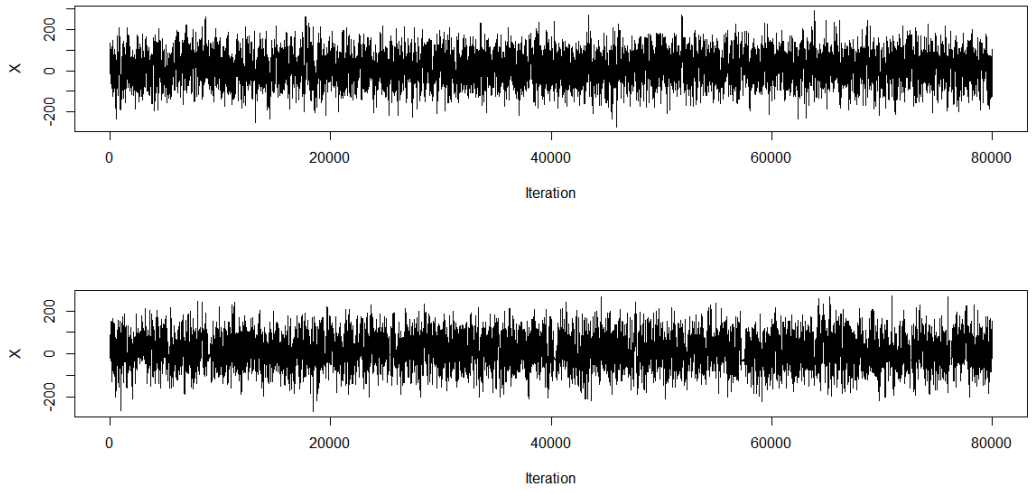

Figure 4.14: Two trace plots showing the respective hot state mixing at inverse temperature $\beta = 0.35^6$ of the runs from the top plotted examples in Figures 4.12 and 4.11, which are for the PT and HAT algorithms respectively. Note that the global mixing for the hot state appears to be better than that of the HAT algorithm when the chain can become temporarily stuck in local modes even in this hot state.

4.5 Computational Expense of the HAT Algorithm

There is no denying the fact that the practical implementation of the HAT algorithm is significantly more expensive per iteration (ignoring inferential quality) over the PT algorithm.

This is due to the expense incurred evaluating the target distribution for the HAT targets. This involves the numerical calculation (and inversion of) of a hessian at **every** evaluation of the target distribution. This is an $O(d^3)$ operation whether or not eigenvalues are calculated to assess whether the hessian is suggestive of a "proper" positive definite covariance matrix.

For comparison, the computational expense of evaluating the target in the five dimensional example in Section 4.4.1 is highlighted with a HAT target evaluation taking typically around 240 times longer than for the toy (very cheap) powered version. Much of the expense can be put down to the chosen tuning of the numerical "hessian" function in R, from the package Gilbert *et al.* [2006], which has been tuned for high accuracy over speed. The hessian calculation computes 25 entries for a 5×5 matrix and for each entry the accuracy tuning parameter was allowed 10 iterations

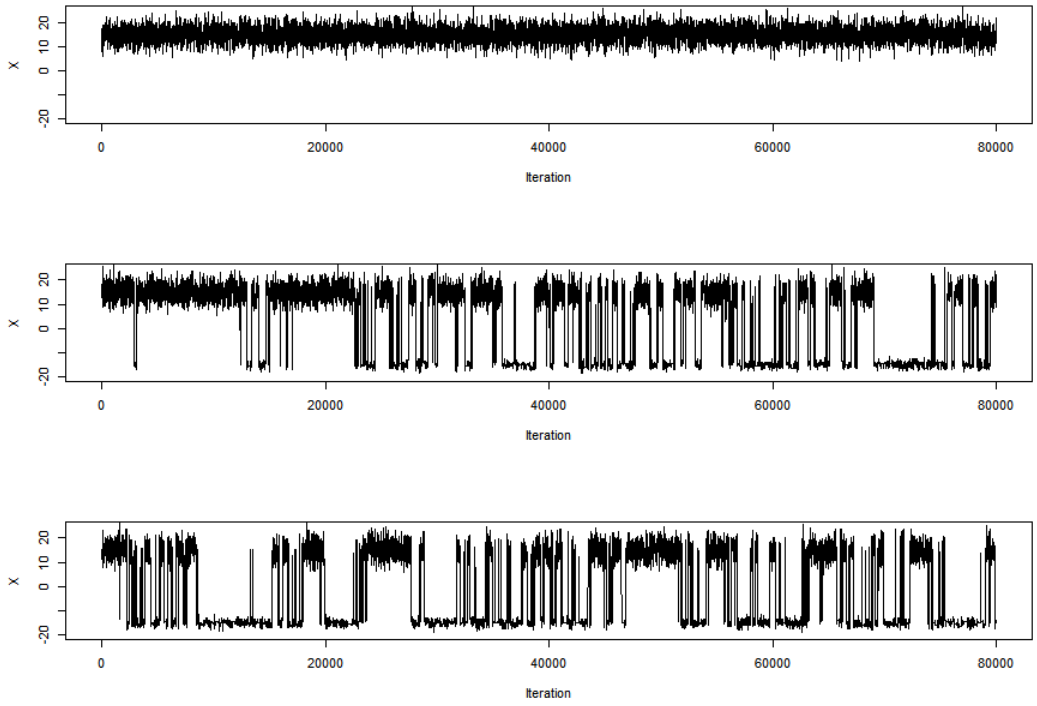

Figure 4.15: Three trace plots of the cold state chain targeting the distribution in equation (4.44) using the PT, HAT and Ideal algorithms respectively. However, unlike the versions of the runs in Figure 4.13, the temperature schedules were on a more ambitious, sub-optimal, spacing. All cold level swap move acceptance rates were approximately 0.16; even for the PT runs.

until convergence; immediately explaining the expense factor of approximately 250.

A key question is how the performance scales with dimension. The cost of forming and inverting a hessian in d-dimensions is $O(d^3)$, which is undeniably expensive.

Particularly in the colder states, with significantly different covariance structures between modes, performing location dependent moves would be essential to ensure fast intra-modal mixing. Position dependent RWM moves, Livingstone [2015], which use the hessian at the current location to estimate the local covariance structure would be one approach to ensure fast intra-modal mixing. It would make the iterative cost $O(d^3)$ for the standard PT algorithm. At each step of the HAT algorithm the hessian can be calculated and stored for use in a position dependent RWM framework making the within temperature moves more efficient at no signif-

icant extra cost.

4.5.1 Limiting Diffusion for the Ideal Algorithm

A key insight into the scalability of HAT comes from collaborative work in conjunction with this thesis by Professor Gareth Roberts (University of Warwick) and Professor Jeffrey Rosenthal (University of Toronto). This work is in the process of being written up for publication at the time of submission of this thesis.

The work analyses a simulated tempering algorithm targeting a bimodal Gaussian target in d-dimensions, with inverse temperature level targets suggested by the ideal algorithm of Section 4.2, i.e.

$$\pi_\beta(x) \propto \sum_{k=1}^{2} w_k \phi_{\left(\mu_k, \frac{\Sigma_k}{\beta}\right)}(x) \tag{4.47}$$

where $\phi_{(\mu,\Sigma)}(.)$ is the density function of a d-dimensional Gaussian with mean μ and variance matrix Σ. The weights of the modes are even with $w_1 = w_2 = 0.5$, $\mu_1 = (-1, \ldots, -1)$, $\mu_2 = (1, \ldots, 1)$, $\Sigma_1 = I_d$ and $\Sigma_2 = \sigma^2 \times I_d$.

Previous analysis in Roberts and Rosenthal [2014], focussed on the ST approach with power-based tempered targets and made the unrealistic assumption that exact, immediate mixing was happening within each temperature level.

The new work analyses the performance of a simulated tempering algorithm where the hot states are given as in equation (4.47) following the idealised target concept of Section 4.2. It makes two, far more realistic assumptions for the mixing of the chain:

1. Immediate mixing within a **single** mode. Hence, conditional on being in one of the mixture components, the Markov chain immediately mixes to invariance within that component.

2. Immediate hot state mixing between modes only at the hot state temperature level.

The first assumption is very realistic, while the second assumption is less so; in particular this will likely be violated for the HAT algorithm (see Section 4.6).

Further to this, the temperature spacings are geometric with $O(d^{-1/2})$ spacings which are suitable and indeed optimal considering the associated optimal scaling results that will follow in Chapter 5. Additionally, the hottest temperature is assumed to be $O(d^{-1})$ to induce stable probabilities of swapping between regions.

The $d+1$ dimensional chain at time t is denoted as (β_t, X_t) where X_t is the location in the state space, $\mathcal{X}$, and β_t is the inverse temperature level. The aim is to find a limiting diffusion for the signed "temperature" component of the chain defined as

$$Y_t = \text{sgn}(X_t)\frac{\log\left(\beta_t/\beta_{\min}\right)}{\log(1/\beta_{\min})} \in [-1, 1]$$

where $\text{sgn}(X_t)$ is 1 if the chain is assigned to the mode centred on $\{1\}^d$ or -1 if the chain is assigned to the mode centred on $\{-1\}^d$ and $\beta_{\min}$ is the minimum of the inverse temperature levels (i.e. hottest state).

With suitable scaling of the process, and using the two assumptions above, it is concluded that Y_t converges to a limiting process characterised as a skew Brownian motion. The scaling that is required to obtain this non-trivial limiting process gives insight into the convergence rate of this particular algorithm as dimensionality grows. It turns out that time must be "sped up" by a factor of $O\left(d\log(d)^2\right)$ to obtain a non-trivial limiting process. This suggests that the convergence time of the algorithm is polynomial in dimension.

This is a positive result for the HAT algorithm since assuming similar behaviour, the added $O(d^3)$ complexity that Hessian information requires suggests that HAT converges in $O(d^4\log(d)^2)$, which is still polynomial in dimension. Comparing this to the standard ST approach for this example, which is torpidly mixing and so convergence is decaying exponentially badly in dimension, see Woodard *et al.* [2009b]. This result is therefore very positive and supportive of the HAT approach.

Alas, there are still open issues with the HAT approach that will likely cause issues with the mixing at the hot temperature. Details are given in the following section.

4.6 Hotter State Within Temperature Proposals

When power tempering, the bottleneck of information transfer through the temperature schedule occurs towards the colder temperatures when there is a relatively very sudden regional weight indifference between consecutive temperature levels. Section 4.4.1 shows that the HAT algorithm can vastly improve the inter-modal mixing and in the canonical setting has comparable iterative performance to the ideal algorithm (albeit at a non trivial computational expense).

However, there is a new bottleneck that if ignored, can hugely undermine the performance of the HAT algorithm. Figure 4.16 shows that the target distributions at the hottest levels essentially becomes a step function along the region boundary

formed by the zones of uncertainty. Such step function targets exhibit poor mixing performance when using the typical hot state style proposals of symmetric RWM.

To see this, consider the bi-modal Gaussian mixture target given in equation (4.4). At an inverse temperature level, β, close to 0, the target distribution at a point associated with the i^{th} mode is proportional to $|\Sigma_i|^{-1/2}$. Then a symmetric RWM move proposed from x to y in this hottest state has (approximately) an acceptance ratio of

$$A \approx \min\left(1, \frac{|\Sigma_{i(y)}|^{-1/2}}{|\Sigma_{i(x)}|^{-1/2}}\right) \tag{4.48}$$

where $i(z) \in \{1, 2\}$ is the modal assignment of position z. Clearly then RWM can mean that the hot states are still "multi-modal" with the chain at the hottest levels trapped in regions with the smallest $|\Sigma_i|$. This is the opposite of the purpose of tempering where the idea is that the chain at the hottest states can move freely about all regions of significant probability mass.

Two solutions to this were initially considered but neither provided performance that exceeded the symmetric RWM dynamics. Both were based on using the ratio of the normalisation constants from the proposal densities to cancel out the ratio of the target distributions (which is very large/small if the jump is inter-regional) in the Metropolis-Hastings acceptance ratio. Suppose that the current location of the Markov chain is $\mathbf{x}$, then these two suggested mechanisms for proposing a new location $\mathbf{y}$ in the hottest states were as follows:

1. Uniform Proposals: Propose component-wise independent symmetric centred uniform proposals such that $y_i \sim \text{Unif}(x_i - a(\mathbf{x}), x_i + a(\mathbf{x}))$ where $a(\mathbf{x}) = \left[2s[\pi_\beta^H(\mathbf{x})]^{1/d}\right]^{-1}$ with s, a tuning parameter. The proposal density is given by

$$q(\mathbf{x}, \mathbf{y}) = \frac{1}{s^d[\pi_\beta^H(\mathbf{x})]} \prod_{i=1}^{d} \mathbb{1}_{|x_i - y_i| \leq a(\mathbf{x})}.$$

   The key is that the acceptance rate for a proposed swap is given by

$$A = \min\left(1, \prod_{i=1}^{d} \mathbb{1}_{|x_i - y_i| \leq h}\right) \tag{4.49}$$

   where $h = \min(a(\mathbf{x}), a(\mathbf{y}))$. Trial runs show that it is very hard to tune this move well and the discontinuous nature of the move becomes a real issue for the harder interesting scenarios where the step jump in the target is large. In fact with any high dimensional setting it is unclear that this move would be able to overcome any "zone of uncertainty".

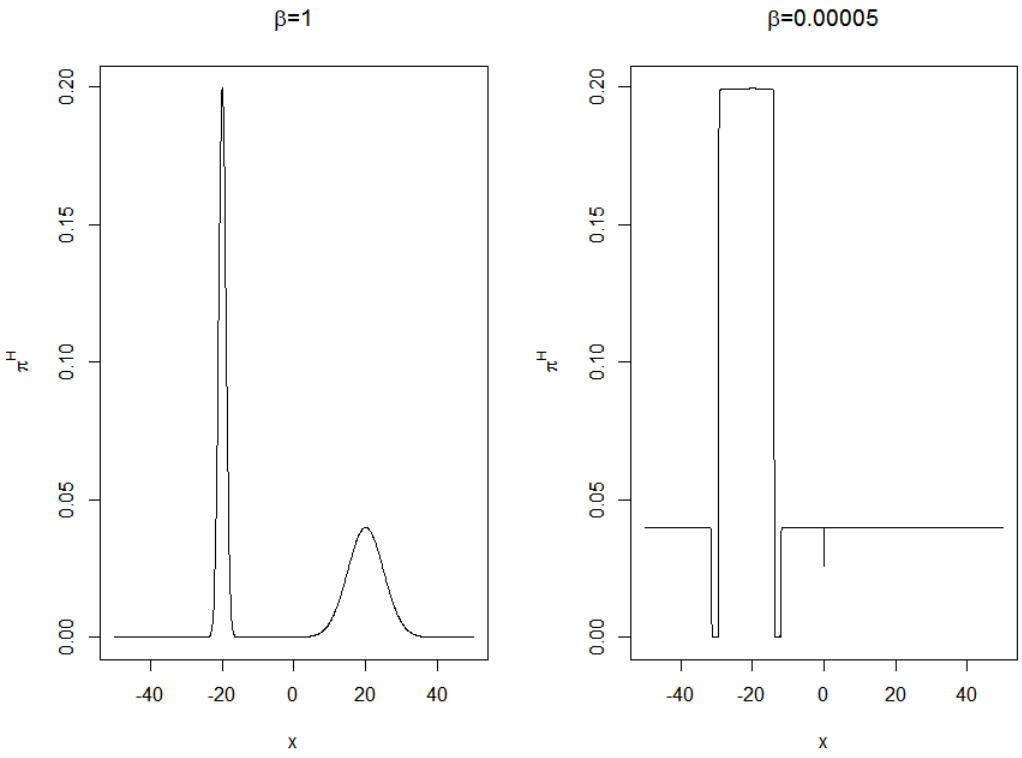

Figure 4.16: Left: Plot of an example bimodal Gaussian mixture. Right: Plot of the HAT target at the inverse temperature level $\beta = 0.00005$. Note the step nature of the function with jumps that would be problematic for a symmetric RWM to overcome.

2. Multivariate-t Proposals: Propose from a Multivariate t-distribution with a suitable scale matrix (given below). It seemed intuitive that the heavy tails of the t-distribution would provide a greater ability for the chain to jump (ambitiously) across the boundary between regions in the hottest states of the HAT algorithm. Hence,

$$\mathbf{y} \sim \text{Multivariate-t}(\nu, \Sigma(x))$$

where ν is the degrees of freedom and Σ_x is the scale matrix of the multivariate-t and in this case is given by a d-dimensional diagonal matrix with the (i,i) entry given by $\Sigma(x)_{(i,i)} = s\left[\pi_\beta^H(x)\right]^{2/d}$ where s is a tuning parameter. Then

the proposal density is given by

$$q_{(\nu,s)}(\mathbf{x},\mathbf{y}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{d/2}|\Sigma(x)|^{1/2}}\left[1+\frac{\nu}{2}(\mathbf{y}-\mathbf{x})'\Sigma(x)^{-1}(\mathbf{y}-\mathbf{x})\right]^{-\frac{\nu+d}{2}}$$

which gives a MH acceptance probability of

$$A = \min\left(1, \left[\frac{1+\frac{\sum_{i=1}^{d}(x_i-y_i)^2}{s\nu\left[\pi_\beta^H(x)\right]^{2/d}}}{1+\frac{\sum_{i=1}^{d}(x_i-y_i)^2}{s\nu\left[\pi_\beta^H(y)\right]^{2/d}}}\right]^{\frac{(v+d)}{2}}\right).$$

It should be noted that the scaling parameter, s introduced above is very important in determining the performance of the algorithm and is directly dependent on the normalisation constant of the target distribution.

Trial runs on hard one dimensional step type targets appear to show that the use of the t-distribution is only as good at traversing the large step down as the symmetric RWM moves with regards to the swap acceptance rate of proposed moves across the step boundary.

A Population Based Approach: Another suggestion that certainly has issues in the higher dimensions is to take a population approach to the target at the hottest level. The motivation for this is that the mixing quality of the chains at the colder temperature relies almost entirely on the hot state ability to traverse the state space fast and efficiently. Therefore, if the mixing in the hot state is "sticky" then this can filter up through to the colder chains when using the HAT algorithm.

If one has a population of particles at the hot state from which one is uniformly selected to undertake a swap move to the consecutively colder temperature then one is not relying upon a single chain.

However, the acceptance rate of a step down can be seen to decay exponentially fast in dimension and so a population would have to grow exponentially in dimension to ultimately negate the step jump problem assuming the particles continue using symmetric RWM.

In low/mid-dimensional problems it is not unrealistic that this could work well and alleviate some of the problem. Since the mixing of the $O(d^{1/2})$ temperature levels depends on the mixing at this hot state it is sensible to put the extra effort in at this single hot temperature to overcome the stickiness issues.

An Implicit Move Approach: Another approach that is being considered as further work to this thesis is the use of implicit proposals which are well moti-

vated in a simple one-dimensional toy step function target but not obviously yet generalisable to a multi-dimensional setting.

New Approach: An interesting idea for a future approach that attempts to preserve regional weight but maintain rapid mixing is suggested in the further work Section 6.2.2. It is an idea that utilises statespace augmentation in an attempt to overcome the core issues with significantly different entropy levels that can occur between modes.

Chapter 5

Optimal Scaling of a Regionally Weight-Preserved PT Algorithm

5.1 Introduction

Atchadé *et al.* [2011] motivated seeking an optimal temperature schedule selection for the efficiency of the transfer of the hot state mixing information through to the cold state. One measure of the transfer efficiency through the temperature schedule is the Expected Squared Jumping Distance ($ESJD_\beta$) for a temperature swap move, which is used as the metric in Atchadé *et al.* [2011] and is given rigorous justification in Roberts and Rosenthal [2014], given by

$$ESJD_\beta = \mathbb{E}\left[(\gamma - \beta)^2\right] \tag{5.1}$$

where β is the current temperature of the chain and γ is the random variable taking the values β if the proposed swap move is rejected or $\beta^{'}$ if the move is accepted.

In order to pass the information efficiently from the hot state to the cold state then one needs a strategy to balance making overly ambitious large jump proposals which have low acceptance probabilities against under ambitious small jump proposals with high acceptance both leading to slow mixing. By tuning the consecutive temperature spacings to maximise the $ESJD_\beta$ between levels then a strategy balancing ambition and acceptance should be reached.

5.1.1 Assumptions and Setup

Suppose that the d-dimensional state space, $\mathcal{X}_d$, can be divided up into distinct regions of significant probability mass, an idea similar to the decompositions used in Woodard *et al.* [2009a]. The aim being that on these regions the distributional mass will be preserved under power-based tempering. Let there be K distinct regions such that

$$\mathcal{X}_d = \cup_{j=1}^{K} A_{(j,d)}. \tag{5.2}$$

This is the case (at least approximately) in the HAT algorithm, especially in the canonical setting with well separated Gaussian modes, where the space is divided up into regions defined by their associated mode.

Assume that there are $n+1$ d-dimensional chains, $\mathbf{x}_0, \ldots, \mathbf{x}_n$, running in parallel at inverse temperature levels, $1 = \beta_0 < \beta_1 < \ldots < \beta_n$ targeting the product distribution

$$\pi_d(\mathbf{x}_0, \ldots, \mathbf{x}_n) \propto \pi_{(w,d)}^{\beta_0}(\mathbf{x}_0) \ldots \pi_{(w,d)}^{\beta_n}(\mathbf{x}_n) \tag{5.3}$$

where the d-dimensional target distribution at inverse temperature level β is taken to be of the weight preserving form

$$\begin{aligned} \pi_{(w,d)}^{\beta}(\mathbf{x}) &\propto \sum_{k=1}^{K} w_k \pi_k^{\beta}(\mathbf{x}) \mathbb{1}_{[\mathbf{x} \in A_{(k,d)}]} \\ &= \sum_{k=1}^{K} w_k \left[\prod_{i=1}^{d} \frac{f_{(k,i)}^{\beta}(x_i)}{\int_{A_k^i} f_{(k,i)}^{\beta}(z) dz} \right] \mathbb{1}_{[\mathbf{x} \in A_{(k,d)}]} \end{aligned} \tag{5.4}$$

where

$$w_k = \left[\int_{A_{(k,d)}} \pi_k(\mathbf{x}) d\mathbf{x} \right] \quad \text{and} \quad \sum_{k=1}^{K} w_k = 1.$$

For simplicity (and tractability of the result), let each region be given by a hypercube

$$A_{(k,d)} = A_k^1 \otimes \ldots \otimes A_k^d = [a_k^1, b_k^1] \otimes \ldots \otimes [a_k^d, b_k^d] \tag{5.5}$$

so that for $i, j \in \{1, \ldots, d\}$, $(b_k^i - a_k^i) = (b_k^j - a_k^j)$.

Furthermore, it is assumed that the unnormalised univariate distributions $f_{(k,i)}^{\beta}(\cdot)$ have a shifted iid form on the corresponding region A_k. That is, for each $k \in \{1, \ldots, K\}$ there is a density, denoted $f_k(\cdot)$, such that for all $i \in \{1, \ldots, d\}$

$$f_{(k,i)}(x_i) = f_k(x_i - \mu_k^i), \tag{5.6}$$

where $\mu_k^i = \frac{a_k^i + b_k^i}{2}$.

The interpretation of the target is that the one-dimensional components are conditionally independent (and shifted identically distributed) given the region. This target preserves the masses in each region throughout the temperature schedule.

In what follows assume that invariance of the chains at all temperature levels has been reached. Suppose that a proposed temperature swap move between the particles $\mathbf{x}$ and $\mathbf{y}$ has been made and that these chains are at the consecutive tempering levels β and $\beta' = \beta + \epsilon$ where $\epsilon = \ell/d^{\frac{1}{2}}$. Thus

$$\mathbf{x} \sim \pi^{\beta}_{(w,d)} \quad \text{and} \quad \mathbf{y} \sim \pi^{\beta'}_{(w,d)}.$$

Then the acceptance probability of the swap move proposed is

$$\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\pi^{\beta'}_{(w,d)}(\mathbf{x})\pi^{\beta}_{(w,d)}(\mathbf{y})}{\pi^{\beta'}_{(w,d)}(\mathbf{y})\pi^{\beta}_{(w,d)}(\mathbf{x})}. \tag{5.7}$$

In the parallel tempering algorithm there are also within temperature mixing type moves that provide the mixing within each temperature level. As in Atchadé *et al.* [2011], the assumption that the chains mix "infinitely" fast within each level relative to the temperature space mixing is made here. Although this is unrealistic at colder temperatures where multimodality prevents effective within temperature mixing it means that the pair of chains, $\mathbf{x}$ and $\mathbf{y}$, at the different levels can be considered independent in the following analysis.

Under the above conditions, the following optimal scaling result will be proved (where $\Phi_{(0,1)}$ is the cumulative distribution function of a standard Gaussian distribution).

Theorem 5.1.1 (Optimal Scaling for a Regionally Weight Preserved Parallel Tempering Algorithm)**.** *Consider the parallel tempering algorithm targeting a distribution defined on a d-dimensional statespace, $\mathcal{X}_d$, which can be decomposed as the union of disjoint hypercubes as described in equations (5.2) and (5.5) and further suppose that the target takes the regionally conditionally iid form given in equations (5.3), (5.4) and (5.6). Then as $d \to \infty$, the ESJD given in equation (5.1) is maximised when ℓ is chosen to maximise*

$$\ell^2 \sum_{j=1}^{K} \sum_{m=1}^{K} 2w_j w_m \Phi_{(0,1)} \left(-\frac{\ell \sigma_{j,m}(\beta)}{2} \right)$$

where $\sigma_{j,m}(\beta) = (I_j(\beta) + I_m(\beta))^{\frac{1}{2}}$ *and* $I_k(\beta) = Var_{f_k^\beta}((\log f_k)(x))$.

Furthermore, the corresponding optimal swap move acceptance rate, $\hat{a}$*, induced between two consecutive temperatures is in the region* $0 < \hat{a} \leq 0.234$ *(3 s.f.).*

The proof of this result is given immediately in the following Section 5.1.2 and is broken down into 3 key steps: establishing limiting Gaussianity of the logged swap move acceptance ratio; computation of the limiting $ESJD_\beta$ and optimisation of the $ESJD_\beta$ and derivation of the corresponding optimal swap acceptance rate. These steps are broken down into three separate lemmas; the results and derivations of Lemmas 5.1.3 and 5.1.4 establish the proof of Theorem 5.1.1.

5.1.2 Proof of Theorem 5.1.1

Recall that the $ESJD_\beta$ takes the form

$$\begin{aligned}
\mathbb{E}_{\pi_d}\left[(\gamma-\beta)^2\right] &= \epsilon^2 \times \mathbb{E}_{\pi_d}\left[\mathbb{P}(\text{accept the swap})\right] \\
&= \epsilon^2 \times \mathbb{E}_{\pi_d}\left[1 \wedge \frac{\pi^{\beta'}_{(w,d)}(\mathbf{x})\pi^{\beta}_{(w,d)}(\mathbf{y})}{\pi^{\beta'}_{(w,d)}(\mathbf{y})\pi^{\beta}_{(w,d)}(\mathbf{x})}\right] \\
&= \epsilon^2 \times \mathbb{E}_{\pi_d}\left[1 \wedge e^B\right], \qquad (5.8)
\end{aligned}$$

where

$$\begin{aligned}
B &= \log\left(\frac{\pi^{\beta'}_{(w,d)}(\mathbf{x})\pi^{\beta}_{(w,d)}(\mathbf{y})}{\pi^{\beta'}_{(w,d)}(\mathbf{y})\pi^{\beta}_{(w,d)}(\mathbf{x})}\right) \\
&= \left[\log\left(\pi^{\beta'}_{(w,d)}(\mathbf{x})\right) - \log\left(\pi^{\beta}_{(w,d)}(\mathbf{x})\right)\right] + \left[\log\left(\pi^{\beta}_{(w,d)}(\mathbf{y})\right) - \log\left(\pi^{\beta'}_{(w,d)}(\mathbf{y})\right)\right] \\
&=: H^{\beta}_d(\mathbf{x}) + H^{\beta'}_d(\mathbf{y}). \qquad (5.9)
\end{aligned}$$

The first step is to understand the asymptotic nature of B. Lemma 5.1.2 establishes this and shows that B has an asymptotically Gaussian mixture distribution.

Lemma 5.1.2 (Asymptotic Gaussianity of B)**.** *Under the setting of Theorem 5.1.1 and* B *given above in equation (5.9); as* $d \to \infty$*,* B *converges weakly to a Gaussian mixture given*

$$\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{1}_{[\mathbf{x}\in A_i]}\mathbb{1}_{[\mathbf{y}\in A_j]} N\left(-\frac{\ell^2}{2}\left(I_j(\beta)+I_m(\beta)\right), \ell^2\left(I_j(\beta)+I_m(\beta)\right)\right), \qquad (5.10)$$

where $I_k(\beta) = Var_{f_k^\beta}\left(\log\left(f_k\right)\right)$.

Proof. Consider the term $H_d^\beta(\mathbf{x})$ from equation (5.9):

$$\begin{aligned}
H_d^\beta(\mathbf{x}) &= \sum_{k=1}^{K}\left[\sum_{i=1}^{d}\log\left(\frac{f_{(k,i)}^{\beta'}(x_i)}{\int_{A_k^i} f_{(k,i)}^{\beta'}(z)dz}\right) - \log\left(\frac{f_{(k,i)}^{\beta}(x_i)}{\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz}\right)\right]\mathbb{1}_{[\mathbf{x}\in A_k]} \\
&= \sum_{k=1}^{K}\left[\sum_{i=1}^{d}\epsilon\log\left(f_{(k,i)}(x_i)\right) - \log\left(\int_{A_k^i} f_{(k,i)}^{\beta'}(z)dz\right) + \log\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right)\right]\mathbb{1}_{[\mathbf{x}\in A_k]} \\
&= \sum_{k=1}^{K}\left[\sum_{i=1}^{d}\epsilon\log\left(f_{(k,i)}(x_i)\right) - \epsilon\frac{\partial}{\partial\beta}\log\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right)\right. \\
&\qquad -\frac{\epsilon^2}{2}\frac{\partial^2}{\partial^2\beta}\log\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right) \\
&\qquad \left.-\frac{\epsilon^3}{6}\frac{\partial^3}{\partial^3\beta}\log\left(\int_{A_k^i} f_{(k,i)}^{[\beta+\xi_{(k,i)}]}(z)dz\right)\right]\mathbb{1}_{[\mathbf{x}\in A_k]} \qquad (5.11)
\end{aligned}$$

where the final line uses a Taylor expansion to third order and $0 < |\xi_{(k,i)}| < \epsilon$ is the mean value Taylor remainder.

Considering just the first derivative term in the final line of equation (5.11)

$$\begin{aligned}
M_{(k,i)}(\beta) := \frac{\partial}{\partial\beta}\log\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right) &= \frac{\int_{A_k^i}\log\left(f_{(k,i)}(z)\right) f_{(k,i)}^{\beta}(z)dz}{\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz} \\
&= \mathbb{E}_{f_{(k,i)}^{\beta}}\left(\log\left(f_{(k,i)}\right)\right) \\
&= \mathbb{E}_{f_k^\beta}\left(\log\left(f_k\right)\right) \qquad (5.12)
\end{aligned}$$

where the final equality is given due to the "shifted" iid form of the target in equation (5.6). Hence, the dependence on the component identifier, i, in the term $M_{(k,i)}(\beta)$ can be dropped and so is instead denoted by $M_k(\beta)$.

Now consider the second order derivative from the final line of equation (5.11)

$$\begin{aligned}
I_{(k,i)}(\beta) &:= \frac{\partial^2}{\partial^2\beta}\log\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right) \\
&= \frac{\int_{A_k^i}\log\left(f_{(k,i)}(z)\right)^2 f_{(k,i)}^{\beta}(z)dz}{\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz} - \frac{\int_{A_k^i}\log\left(f_{(k,i)}(z)\right) f_{(k,i)}^{\beta}(z)dz}{\left(\int_{A_k^i} f_{(k,i)}^{\beta}(z)dz\right)^2} \\
&= \mathrm{Var}_{f_{(k,i)}^{\beta}}\left(\log\left(f_{(k,i)}\right)\right) \\
&= \mathrm{Var}_{f_k^\beta}\left(\log\left(f_k\right)\right) \qquad (5.13)
\end{aligned}$$

where the final equality is given due to the assumed "shifted" iid setting described above in equation (5.6). Hence, the dependence on the component identifier, i, in the term $I_{(k,i)}(\beta)$ can be dropped and so is instead denoted by $I_k(\beta)$.

For notational convenience, and due to the "shifted" iid setup, herein the following notation is used:

$$J_k(\beta) := \frac{\partial^3}{\partial^3 \beta} \log \left(\int_{A_k^i} f_{(k,i)}^{\beta}(z) dz \right) \qquad \forall i \in \{1, \ldots, d\}$$

and furthermore the Taylor remainder from (5.11) can be given as ξ_k without a dependence on i.

Using this new notation; the shifted iid form of the conditional components given in equation (5.6) and writing $x_i^s = x_i - \mu_k^i$ then $H_d^{\beta}(\mathbf{x})$ can be rewritten as

$$H_d^{\beta}(\mathbf{x}) = \sum_{k=1}^{K} \left[\sum_{i=1}^{d} \epsilon \log \left(f_k(x_i^s) \right) - \epsilon M_k(\beta) - \frac{\epsilon^2}{2} I_k(\beta) - \frac{\epsilon^3}{6} J_k(\beta + \xi_k) \right] \mathbb{1}_{[\mathbf{x} \in A_k]}.$$

By identical methodology to computing $H_d^{\beta}(\mathbf{x})$ and with $y_i^s = y_i - \mu_k^i$ then

$$H_d^{\beta'}(\mathbf{y}) = \sum_{k=1}^{K} \left[\sum_{i=1}^{d} -\epsilon \log \left(f_k(y_i^s) \right) + \epsilon M_k(\beta) + \frac{\epsilon^2}{2} I_k(\beta) + \frac{\epsilon^3}{6} J_k(\beta + \xi_k) \right] \mathbb{1}_{[\mathbf{y} \in A_k]}.$$

However, one can write

$$M_k(\beta') = M_k(\beta) + \epsilon I_k(\beta) + \frac{\epsilon^2}{2} J_k(\beta + \xi_{T_k})$$

where ξ_{T_k} is the the Taylor correction term such that $0 < |\xi_{T_k}| < \epsilon$. Substituting this term for $M_k(\beta)$ into the above expression for $H_d^{\beta'}(\mathbf{y})$,

$$\begin{aligned} H_d^{\beta'}(\mathbf{y}) \quad = \quad & \sum_{k=1}^{K} \left[\sum_{i=1}^{d} -\epsilon \log \left(f_k(y_i^s) \right) + \epsilon M_k(\beta') - \frac{\epsilon^2}{2} I_k(\beta) \right. \\ & \left. + \frac{\epsilon^3}{6} \left(J_k(\beta + \xi_k) - 3 J_k(\beta + \xi_{T_k}) \right) \right] \mathbb{1}_{[\mathbf{y} \in A_k]}. \end{aligned}$$

Let $E_k^{\mathbf{x}}$ denote the event that $\mathbf{x} \in A_k$ then, with a slight abuse of notation, condi-

tioning B on the events $E_l^{\mathbf{x}}$ and $E_m^{\mathbf{y}}$

$$\begin{aligned} B|E_j^{\mathbf{x}}, E_m^{\mathbf{y}} &= \sum_{i=1}^{d} \left[\epsilon \left(\log \left(f_j(x_i^s) \right) - M_j(\beta) \right) - \epsilon \left(\log \left(f_m(y_i^s) \right) - M_m(\beta^{'}) \right) \right. \\ &\quad \left. - \frac{\epsilon^2}{2} \left(I_j(\beta) + I_m(\beta) \right) - \frac{\epsilon^3}{2} J_k(\beta + \xi_{T_k}) \right]. \end{aligned}$$

Defining

$$\begin{aligned} R_{(x,y,j,m)}^{(\beta,\beta^{'})} &:= \sum_{i=1}^{d} \left[\epsilon \left(\log \left(f_j(x_i^s) \right) - M_j(\beta) \right) - \epsilon \left(\log \left(f_m(y_i^s) \right) - M_m(\beta^{'}) \right) \right. \\ &\quad \left. - \frac{\epsilon^2}{2} \left(I_j(\beta) + I_m(\beta) \right) \right] \\ &=: \sum_{i=1}^{d} r_{(x,y,j,m),i}^{(\beta,\beta^{'})} \end{aligned} \tag{5.14}$$

then, conditional on the events $E_l^{\mathbf{x}}$ and $E_m^{\mathbf{y}}$, the $r_{(x,y,l,m),i}^{(\beta,\beta^{'})}$ are independent and identically distributed for all $i \in \{1, \ldots, d\}$. Using the assumed independence between the $\mathbf{x}$ and $\mathbf{y}$ then

$$\mathbb{E}_{\pi_d} \left(r_{(x,y,j,m),i}^{(\beta,\beta^{'})} \right) = -\frac{\epsilon^2}{2} \left(I_j(\beta) + I_m(\beta) \right) \tag{5.15}$$

and

$$\begin{aligned} \mathrm{Var}_{\pi_d} \left(r_{(x,y,j,m),i}^{(\beta,\beta^{'})} \right) &= \epsilon^2 \left(\mathrm{Var}_{f_j^{\beta}} \left(\log \left(f_j \right) \right) + \mathrm{Var}_{f_m^{\beta^{'}}} \left(\log \left(f_m \right) \right) \right) \\ &= \epsilon^2 \left(I_j(\beta) + I_m(\beta) \right). \end{aligned} \tag{5.16}$$

With $\epsilon = \ell / d^{1/2}$, then using equations (5.16), (5.15), and the central limit theorem for iid random variables, e.g. Durrett [2010], then as $d \to \infty$

$$R_{(x,y,j,m)}^{(\beta,\beta^{'})} \Rightarrow N \left(-\frac{\ell^2}{2} \left(I_j(\beta) + I_m(\beta) \right), \ell^2 \left(I_j(\beta) + I_m(\beta) \right) \right). \tag{5.17}$$

Furthermore, assuming continuity of $J_k(\cdot)$ for all $k = 1, , K$ then

$$\lim_{d \to \infty} J_k(\beta + \xi_{T_k}) = J_k(\beta)$$

and so there exists a bounding constant $C \in \mathbb{R}$ such that for all $k \in \{1, \ldots, K\}$

$$|J_k(\beta + \xi_{T_k})| < C.$$

Consequently,

$$\left| \epsilon^3 \sum_{i=1}^{d} J_k(\beta + \xi_{T_k}) \right| \leq \frac{\ell^3}{d^{1/2}} C \to 0 \text{ as } d \to \infty.$$

Thus, by trivial use of Slutsky's Theorem

$$B|E_j^{\mathbf{x}}, E_m^{\mathbf{y}} \Rightarrow N\left(-\frac{\ell^2}{2} \left(I_j(\beta) + I_m(\beta) \right), \ell^2 \left(I_j(\beta) + I_m(\beta) \right) \right). \tag{5.18}$$

Removing the conditioning across all regions gives the result in Lemma 5.1.2 and the proof is complete. □

Using the result in Lemma 5.1.2 then an expression for the $ESJD_\beta$ for large d is now derived.

Lemma 5.1.3 (Asymptotic $ESJD_\beta$)**.** *Under the setting of Theorem 5.1.1 and B given above in equation (5.9); then for large d the $ESJD_\beta$ is approximately given as*

$$ESJD_\beta = \epsilon^2 \times \mathbb{E}_{\pi_d}\left[1 \wedge e^B\right] \sim \frac{\ell^2}{d} \sum_{j=1}^{K} \sum_{m=1}^{K} w_j w_m \left(2\Phi_{(0,1)}\left(-\frac{\ell \sigma_{j,m}(\beta)}{2} \right) \right), \tag{5.19}$$

where $\sigma_{j,m}(\beta) = \left(I_j(\beta) + I_m(\beta)\right)^{\frac{1}{2}}$.

Proof. Firstly, an essential subsidiary result is established. Recall that (for some general σ) if $G \sim N(-\frac{\sigma^2}{2}, \sigma^2)$ then, as was explicitly derived in equation (3.47), but also found in Roberts *et al.* [1997],

$$\mathbb{E}\left(1 \wedge e^G\right) = 2\Phi_{(0,1)}\left(-\frac{\sigma}{2}\right). \tag{5.20}$$

Recall the form of the $ESJD_\beta$ in this case:

$$ESJD_\beta = \epsilon^2 \times \mathbb{E}_{\pi_d}\left[1 \wedge e^B\right].$$

For large d, B will be assumed to be distributed as in its asymptotic conditional Gaussian form from Lemma 5.1.2. Then, taking the expectation under this assump-

tion and applying the result from equation (5.20) then for large d

$$ESJD_\beta \sim \frac{\ell^2}{d} \sum_{j=1}^{K} \sum_{m=1}^{K} w_j w_m \left(2\Phi_{(0,1)} \left(-\frac{\ell \sigma_{j,m}(\beta)}{2} \right) \right), \tag{5.21}$$

where $\sigma_{j,m}(\beta) = (I_j(\beta) + I_m(\beta))^{\frac{1}{2}}$.

□

Lemma 5.1.3 establishes the key asymptotic form of the $ESJD_\beta$ for Theorem 5.1.1. However, the final part of the theorem statement regarding the optimal spacing and associated optimal acceptance rate is yet to be established.

To this end the aim is to find the optimal spacing and to do this the $ESJD_\beta$ in equation (5.19) must be maximised with respect to ℓ. This can be done numerically to give a value $\hat{\ell}$.

For intuition, the $ESJD_\beta$ function in equation (5.19) is considered for a simple two region example, so $\mathcal{X}_d = \cup_{j=1}^{2} A_{(j,d)}$, where $I_1(\beta) = 1$ and $I_2(\beta) = 5$. Suppose that the weight in each region is equal and so $w_1 = w_2 = 0.5$. Figure 5.1 shows the plot of the $ESJD_\beta$ function from equation (5.19) over a range of values of the scaling parameter, ℓ. It is clear that there is indeed an optimal scaling that finds a balance between over ambitious moves and under-ambitious moves in the temperature schedule.

Optimisation for a specific problem through numerical techniques would require that one knows the values of the $\sigma_{j,m}$ terms. This is not generally the case, and although one would still be able to do trial runs of the algorithm to tune the estimated $ESJD_\beta$ to be maximal, without knowledge of how big this can be then it is hard and computationally expensive to blindly tune to an unknown optimum. In the optimal scaling of the temperature schedule in Atchadé *et al.* [2011], there was an associated optimal acceptance rate for swap moves between consecutive temperature levels and this gives the practitioner a clear tuning target for the spacings.

In this case, the optimal acceptance rate is less clearly attainable due to the summation of terms in equation (5.19) meaning that there won't be a single optimal acceptance rate, see below for the example in Figure 5.2. In Sherlock [2006] a similar problem arises when finding the associated optimal acceptance rate for the tuning of a RWM algorithm on spherically symmetric target distributions. In that case it turned out that through a clever application of Jensen's inequality then the optimal acceptance rate is less than or equal to the typical 0.234 value.

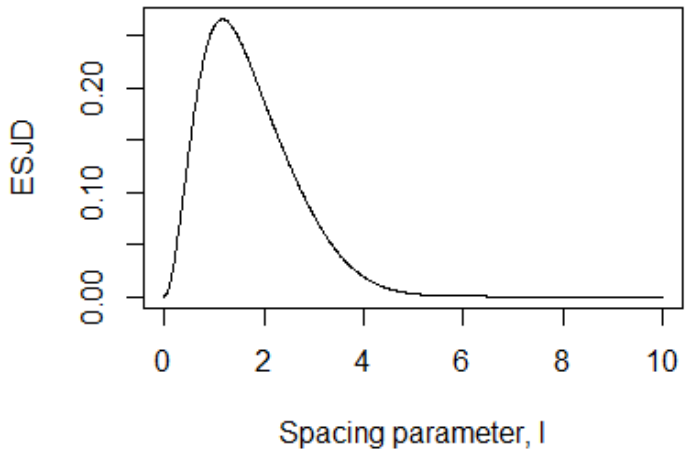

Figure 5.1: Plotting the $ESJD_\beta$ over different values of ℓ in a basic example where $I_1(\beta) = 1$ and $I_2(\beta) = 5$ and $w1 = w2 = 0.5$.

Note that the $ESJD_\beta$ given in equation (5.19) can be expressed as

$$ESJD_\beta \sim \frac{\ell^2}{d}\mathbb{E}_{\Sigma_\beta}\left(2\Phi_{(0,1)}\left(-\frac{\ell\Sigma_\beta}{2}\right)\right)$$

where Σ_β is a discrete RV such that $\mathbb{P}(\Sigma_\beta = \sigma_{i,j}(\beta)) = w_i w_j$. This has similarities to the form of the $ESJD$ in Sherlock [2006] and motivates looking for a similar optimal acceptance rate scaling range.

Recall the example given in Figure 5.1; instead of fixing the weights in the two regions, consider finding the optimal spacings over a range of different values of the weights such that $w_1 = 1 - w_2$. Having found the corresponding optimal spacings then the associated optimal acceptance rates are calculated. Figure 5.2 shows how the optimal acceptance rate changes as the weight of region 1, w_1, moves between the extreme points where there is all or none of the mass in the first region.

At the extreme points the scalings (reassuringly) correspond to the setting of the work in Atchadé *et al.* [2011] and give a corresponding 0.234 optimal scaling. Figure 5.2 shows that in this basic example the optimal acceptance rates lie in the conjectured $\hat{a} \leq 0.234$ region.

Figure 5.2 motivates the existence of an optimal acceptance rate, which one

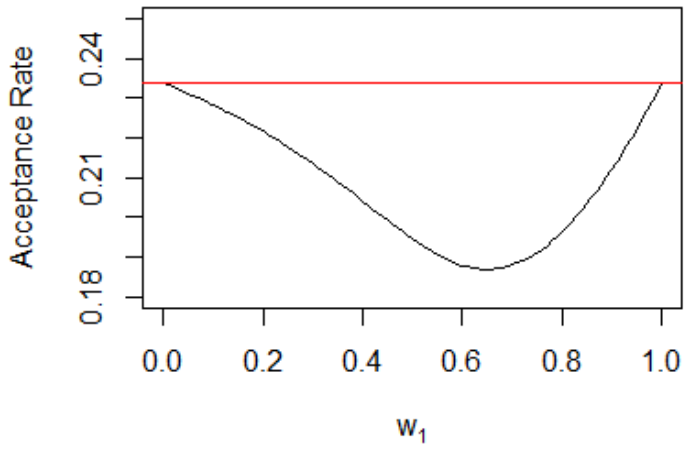

Figure 5.2: Black line: the numerically obtained optimal acceptance rates over different values of w_1 in the basic 2 region example where $I_1(\beta) = 1$ and $I_2(\beta) = 5$. Red line: the 0.234 level.

expects to lie in the range $0 < \hat{a} \leq 0.234$. This will now be derived explicitly.

Lemma 5.1.4 (Optimal Acceptance Rate)**.** *Under the setting of Theorem 5.1.1 then Lemma 5.1.3 showed that for large d the $ESJD_\beta$*

$$ESJD_\beta \sim \frac{\ell^2}{d} \sum_{j=1}^{K} \sum_{m=1}^{K} w_j w_m \left(2\Phi_{(0,1)} \left(-\frac{\ell \sigma_{j,m}(\beta)}{2} \right) \right), \tag{5.22}$$

where $\sigma_{j,m}(\beta) = (I_j(\beta) + I_m(\beta))^{\frac{1}{2}}$. Optimising this with respect to ℓ corresponds to a consecutive temperature level spacing that has an associated optimal temperature swap acceptance rate of $0 < \hat{a} \leq 0.234$.

Proof. For convenience, the $ESJD_\beta$ term established in Lemma 5.1.3 can be reformulated so that the sum with indicators is replaced by a random variable denoted Σ_β. Hence, for large d the $ESJD_\beta$ is (approximately)

$$ESJD_\beta = \frac{\ell^2}{d} \mathbb{E}_{\Sigma_\beta} \left(2\Phi_{(0,1)} \left(-\frac{\ell \Sigma_\beta}{2} \right) \right)$$

where Σ_β is a discrete RV such that $\mathbb{P}(\Sigma_\beta = \sigma_{i,j}(\beta)) = w_i w_j$.

In order to find the optimal spacing, $\hat{\ell}$, this is differentiated to ℓ and set equal to 0. Hence,

$$\frac{\partial}{\partial l} ESJD_\beta = \mathbb{E}_{\Sigma_\beta}\left(4\ell\Phi_{(0,1)}\left(-\frac{\ell\Sigma_\beta}{2}\right) - \ell^2\Sigma_\beta\phi_{(0,1)}\left(-\frac{\ell\Sigma_\beta}{2}\right)\right),$$

and setting this equal to 0 to obtain a formula for the optimal spacing $\hat{\ell}$ gives

$$2\mathbb{E}_{\Sigma_\beta}\left(\Phi_{(0,1)}\left(-\frac{\hat{\ell}\Sigma_\beta}{2}\right)\right) = \mathbb{E}_{\Sigma_\beta}\left(\frac{\hat{\ell}^2\Sigma_\beta}{2}\phi_{(0,1)}\left(-\frac{\hat{\ell}\Sigma_\beta}{2}\right)\right). \tag{5.23}$$

Note that differentiating the $ESJD_\beta$ a second time verifies that the value $\hat{\ell}$ is indeed a maximum. Next, as in Sherlock [2006], define the function $h(\cdot)$ defined so that

$$h(x) = -\Phi^{-1}_{(0,1)}(x)\,\phi_{(0,1)}\left(\Phi^{-1}_{(0,1)}(x)\right) \tag{5.24}$$

and note that $h(\cdot)$ is a concave function since for any $x \in (0,1)$

$$\frac{\partial^2 h}{\partial x^2} = -2\frac{\Phi^{-1}_{(0,1)}(x)}{\phi_{(0,1)}\left(\Phi^{-1}_{(0,1)}(x)\right)} < 0.$$

By considering the form of the $ESJD_\beta$ given in equation (5.8) then it is clear that for any spacing ℓ the associated acceptance rate is given by

$$a = \mathbb{E}_{\Sigma_\beta}\left(2\Phi_{(0,1)}\left(-\frac{\ell\Sigma_\beta}{2}\right)\right)$$

and so the optimal acceptance rate, $\hat{a}$ is given by

$$\hat{a} = \mathbb{E}_{\Sigma_\beta}\left(2\Phi_{(0,1)}\left(-\frac{\hat{\ell}\Sigma_\beta}{2}\right)\right).$$

Letting $V := \Phi_{(0,1)}\left(-\frac{\hat{\ell}\Sigma_\beta}{2}\right)$ then by equation (5.23) at the optimal spacing

$$\hat{a} = 2\mathbb{E}_{\Sigma_\beta}(V) = \mathbb{E}_{\Sigma_\beta}(h(V))$$

where $h(\cdot)$ is given above in equation (5.24).

Since $h(\cdot)$ is concave then Jensen's inequality can be applied to give

$$\mathbb{E}_{\Sigma_\beta}(h(V)) \leq h\left(\mathbb{E}_{\Sigma_\beta}(V)\right)$$

and thus

$$\hat{a} = 2\mathbb{E}_{\Sigma_\beta}(V) \leq -\Phi_{(0,1)}^{-1}\left(\mathbb{E}_{\Sigma_\beta}(V)\right)\phi_{(0,1)}\left(\Phi_{(0,1)}^{-1}\left(\mathbb{E}_{\Sigma_\beta}(V)\right)\right). \tag{5.25}$$

Now let $m := -\Phi_{(0,1)}^{-1}\left(\mathbb{E}_{\Sigma_\beta}(V)\right)$ then by equation (5.25)

$$2\Phi_{(0,1)}(-m) \leq m\phi_{(0,1)}(-m), \tag{5.26}$$

with equality only in the case when m is the optimiser of the function $m^2\Phi_{(0,1)}(-m)$. Let this optimal m be denoted $\hat{m}$ then

$$2\Phi_{(0,1)}(-\hat{m}) = \hat{m}\phi_{(0,1)}(-\hat{m}) = 0.234 \quad (3 \text{ s.f.}). \tag{5.27}$$

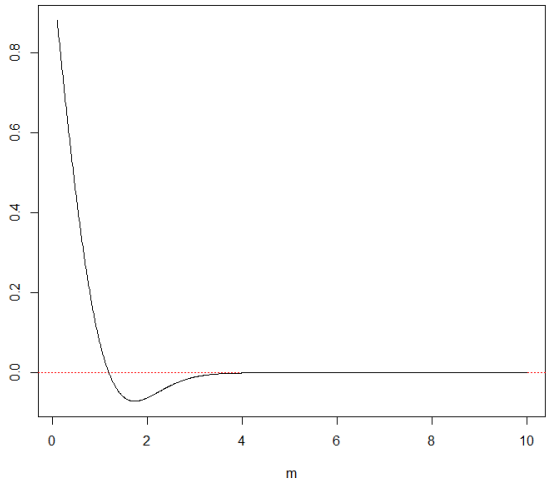

Figure 5.3: Plot of the function $2\Phi_{(0,1)}(-m) - m\phi_{(0,1)}(-m)$

Figure 5.3 shows a plot of the function $2\Phi_{(0,1)}(-m) - m\phi_{(0,1)}(-m)$. Although not entirely clear in the figure, there is only the one point $\hat{m}$ giving a root of the function (occuring in the interval of $[0, 2]$) and then for $m > \hat{m}$ the inequality given above in equation (5.26) holds strictly. Consequently, given that equation (5.26) holds then it implies that in this case $m > \hat{m}$ and so crucially due to the decreasing monotonicity of $\Phi_{(0,1)}(-m)$ the acceptance rate of the algorithm satisfies

$$\hat{a} = 2\Phi_{(0,1)}(-m) \leq 2\Phi_{(0,1)}(-\hat{m}) = 0.234 \quad (3 \text{ s.f.}), \tag{5.28}$$

where the last approximate equality is given from equation (5.27). □

Thus combining the results of Lemmas 5.1.3 and 5.1.4 completes the proof of Theorem 5.1.1.

5.2 Implications and Suggestions of this Optimal Scaling Result

The optimal scaling result in Theorem 5.1.1 gives the practitioner guidance for optimally tuning a regionally weight preserved PT algorithm. It does this by suggesting a range of acceptance rate values (≤ 0.234) that the optimal temperature swap acceptance rate should fall within. Unlike the optimal tuning suggested in Atchadé *et al.* [2011], the theorem only gives a range rather than a fixed value.

As was noted in Atchadé *et al.* [2011], much of the previous literature was suggestive of a geometrically defined temperature schedule. Recalling Section 1.5.2, which reviewed this idea in the context of the optimality result of Atchadé *et al.* [2011]; a geometric schedule would only be optimal in their setting when

$$I(\beta) = \mathrm{Var}_{f^\beta}(\log f) \propto 1/\beta^2.$$

Indeed, this is the case when the target distribution is a uni-modal Gaussian.

However, moving to a multimodal Gaussian mixture then this no longer holds and tuning acceptance rates in a PT algorithm to be 0.234 can be misleading through finite runs of the algorithm (especially in higher dimensions).

Suppose that the HAT algorithm is being used to target a multimodal Gaussian mixture in d-dimensions, then Theorem 5.1.1 gives useful guidance for the setup of the temperature spacings particularly in the colder part of the temperature schedule when indeed there is (at least approximately) an optimal geometric schedule.

Corollary 5.2.1. *Suppose that a target Gaussian mixture distribution is constructed from K mixture components defined so that*

$$\pi_\beta(x) \propto \sum_{k=1}^{K} w_k \phi_{\left(\mu_k, \frac{\Sigma_k}{\beta}\right)}(x)$$

where $\Sigma_i := \sigma_i^2 I_d$. Suppose that all components are well spaced into K separate hypercube regions as in the setup of Theorem 5.1.1; with the μ_i defining the centring vector of the hypercubes.

Assume that this is being targeted by using a regionally weight preserving tempering algorithm. Provided the dimensionality, d, is large then the $ESJD_\beta$ is (approximately) given by

$$ESJD_\beta \approx \frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell}{\beta\sqrt{2}}\right).$$

Maximising this form for the $ESJD_\beta$ derives an associated optimal acceptance rate of 0.234 (3s.f.). Furthermore $\hat{\ell} \propto \beta$ and so the optimal spacing is (approximately) geometric.

Proof. Recall the setting of Theorem 5.1.1; for large d, the $ESJD_\beta$ is given by

$$ESJD_\beta = \frac{\ell^2}{d}\sum_{j=1}^{K}\sum_{m=1}^{K} w_j w_m \left(2\Phi_{(0,1)}\left(-\frac{\ell\sigma_{j,m}(\beta)}{2}\right)\right),$$

where $\sigma_{j,m}(\beta) = (I_j(\beta) + I_m(\beta))^{\frac{1}{2}}$. However, particularly in the cold states, assuming the regional truncations of tails only remove negligible mass then $\forall j, m \in \{1, \ldots, K\}$ then $\sigma_{j,m}(\beta) \approx \left(\frac{2}{\beta^2}\right)^{1/2}$ which is independent of region. Hence the $ESJD_\beta$ term simplifies to

$$ESJD_\beta \approx \frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell}{\beta\sqrt{2}}\right).$$

This has a maximum with respect to ℓ such that $\hat{\ell} \propto \beta$ and with this optimal spacing the induced optimal acceptance rate is directly derived to be

$$ACC = 2\Phi_{(0,1)}\left(-\frac{\hat{\ell}}{\beta\sqrt{2}}\right) = 0.234 \qquad (3s.f.).$$

□

Consequently, a practitioner using a regionally weight preserving PT algorithm targeting a well spaced Gaussian mixture distribution (or something approximately of this form) is encouraged to tune the consecutive temperature swap acceptance rates to an approximately optimal value of 0.234.

5.2.1 The Problem with $ESJD_\beta$

An important question is whether the $ESJD_\beta$ metric of mixing speed throughout the temperature schedule really is the optimal metric to assess the performance

of a regionally weight preserved tempering algorithm. Using standard power-based tempering, optimising the $ESJD_\beta$ is not necessarily inductive of an algorithm giving good inter-modal mixing.

In fact, due to the lack of weight preservation and finite nature of runs of the algorithm, the tuning stategy suggested can result in spacing tunings that are only optimised for temperature swap moves occurring within a subset of the space; potentially leading to critically poor scalings for swap moves elsewhere in the state space.

The optimal temperature schedule scaling results are all under the assumption of infinitely fast mixing within each temperature, which is essentially only realistic if regional weight preservation is satisfied; something certainly not true when using standard power-based tempering. This is major motivation for analysing the joint process of within and temperature moves, as has been done in the new work described in Section 4.5.1.

Indeed, in all the optimal scaling results discussed, the temperature spacings necessarily take the form $\epsilon = \ell/d^{1/2}$ for non-degeneracy of the temperature swap acceptance probabilities. When using standard power-based tempering, it was shown in a heuristic in Section 4.3.1, that the regional weight inconsistency between temperature levels degenerates critically unless the spacings are chosen to be $O(d^{-1})$. This suggests that for the power-based tempering setup, optimising the $ESJD_\beta$ as in Atchadé *et al.* [2011], cannot alone induce a robust algorithm setup for an asymmetric modal problem. Having regional weight preservation overcomes this contradiction of scalings, giving justification to the use of $ESJD_\beta$ in that setting.

There is still a major issue with using $ESJD_\beta$ in **all** scaling results discussed in this thesis. This is a metric in the temperature space and essentially integrates out the dependence on the current location of the chains in the respective consecutive temperature levels. As such, **optimisation of the $ESJD_\beta$ suggests an algorithm that has the fastest mixing speed throughout the temperature schedule with the optimal consecutive spacings tuned in favour of those modes that have the most mass along with the "easiest" swap moves between each level.** This by no means guarantees the best algorithm to enhance the inter-modal mixing of the chains.

To illustrate this consider a bi-modal target in the setting of Theorem 5.1.1. Suppose that $\mathcal{X}_d = A_{(1,d)} \cup A_{(2,d)}$ and that at a given inverse temperature level β then $I_1(\beta) = 1$ and that $I_2(\beta) = V$ for a range of $V \in \{4, 8, 12, 16\}$. Figure 5.4 shows how the optimal acceptance rate and corresponding optimal value of the spacing parameter $\hat{\ell}$, induced by maximising the $ESJD_\beta$ given in Theorem 5.1.1, vary as

the weight assigned to mode 1 varies between 0 and 1 for each of the different values of V.

Figure 5.4 shows that as the difference between the terms $I_1(\beta)$ and $I_2(\beta)$ becomes larger then the unsuitability of the temperature spacing for a given mode becomes evermore inappropriate. Indeed, modes with small $I(\beta)$ are suggestive of a more ambitious spacings.

Consequently, the $ESJD_\beta$, favours the dominant modes which admit the more ambitious spacings due to their having small $I(\beta)$. This is particularly highlighted in Figure 5.4 in the final row of plots when $I_2(\beta) = 16$ where the weight of the first mode, with $I_1(\beta) = 1$, has to drop below roughly 0.35 before the spacing stops being almost entirely tuned to swap moves of chains within the first mode.

Further to this, it is interesting that in this final row of plots there appears to be a jump discontinuity in each plot. It would appear that once the difference between the $I(\beta)$ terms in the two modes becomes large enough the optimal approach with regards to maximising the $ESJD_\beta$ in Theorem 5.1.1 is to (essentially) optimise for swap moves in only one region. Hence the optimally tuned spacing doesn't find a "compromise value" but instead optimises to a single region. Clearly this cannot be the right thing to do if one is trying to optimise the inter-modal mixing since the algorithm will be only tuned to perform well for swap moves of chains that are both within this chosen optimal region.

This evidence suggests that even once weight is preserved regionally and the spacings tuned according to Theorem 5.1.1, the $ESJD_\beta$ metric has not produced an algorithm with optimal inter-modal mixing.

There are two approaches that are intuitively sensible to try to overcome this problem. Their practicality is problematic and implementation is left as further work but nevertheless the concepts are worthy of discussion.

1. Scale the spacings to be optimal with respect to only the mode with the largest value of $I(\beta)$. Hence the optimal spacing would be tuned so that the spacing parameter, ℓ, maximises

$$ESJD \approx \frac{2\ell^2}{d}\Phi_{(0,1)}\left(-\frac{\ell\sqrt{\sigma^2}}{\sqrt{2}}\right)$$

   where $\sigma^2 = \inf_j\{I_j(\beta)\}$.

   This would scale the spacings so that for the majority of modes the spacings are under-ambitious. However, the intuition is that at least there will be no mode that would find a swap move impossible by being far too over-ambitious.

In practice prior knowledge of $I(\beta)$ in each mode is unrealistic. However, if there was suitable regional structure, which is given in the HAT algorithm setting, then monitoring swap moves between chains in the same mode and tuning so that every mode has at least a 0.234 swap acceptance rate would be doing exactly as suggested. This would of course not be easy to implement.

2. Only monitor swap moves that are useful (in the context given below). Although discussed in detail below. The key idea is that the $ESJD_\beta$ considered in Theorem 5.1.1 integrates out the locations of the chains within the temperature levels meaning that swap moves of chains that are in the same region/mode are "averaged out" over also. Such swap moves are useless for aiding inter-modal mixing since no "new information" from the hot state is being transferred to the cold state. An idea is to scale the spacings to have an optimal spacing but only with respect to useful swap type moves.

Pursuing suggestion 2, a non-rigorous definition of a useful swap move is given:

Definition 5.2.1. Consider the parallel tempering algorithm targeting a distribution defined on a d-dimensional statespace, $\mathcal{X}_d$, which can be decomposed as the union of disjoint hypercubes as described in equations (5.2) and (5.5) and further suppose that the target takes the regionally conditionally iid form given in equations (5.3), (5.4) and (5.6). Then define a "useful swap move" to be a swap move that is between chains at consecutive temperature levels that are in different regions.

If one scales the temperature spacings in a regionally weight preserving tempering algorithm to be optimal according to $ESJD_\beta$ under the assumption that the moves are conditionally only useful swap type moves then one is looking to optimise

$$ESJD_\beta^F = \mathbb{E}\left[(\gamma - \beta)^2 \middle| \mathrm{F}\right], \tag{5.29}$$

where F is the event that the chains being swapped are in two different regions of the state space. The regional decomposition is assumed to be such that each region has a single mode; similar to the setup given in equation (5.2).

The following corollary to Theorem 5.1.1, shows that defining the conditioned $ESJD_\beta^F$ as in equation (5.29) and optimising this with respect to the spacing parameter ℓ gives a well defined limiting $ESJD_\beta^F$ and an associated optimal acceptance rate.

Corollary 5.2.2. *Assume the setup from Theorem 5.1.1 with a target distribution that has a regionally conditionally iid form given in equations (5.3), (5.4) and (5.6). Then as $d \to \infty$, the $ESJD^F_\beta$ given in equation (5.29) is optimised when ℓ is chosen to maximise*

$$\ell^2 \sum_{j=1}^{K} \sum_{m \neq j} 2W_{(j,m)} \Phi_{(0,1)} \left(-\frac{\ell \sigma_{j,m}(\beta)}{2} \right)$$

where $\sigma_{j,m}(\beta) = (I_j(\beta) + I_m(\beta))^{\frac{1}{2}}$ and $I_k(\beta) = Var_{f_k^\beta}(\log(f_k))$ and the weights are given by

$$W_{(j,m)} = \frac{w_j w_m}{\sum_{j=1}^{K} \sum_{m \neq j} w_j w_m}.$$

Furthermore, the corresponding optimal swap move acceptance rate, denoted $\hat{a}$, induced between two consecutive temperatures is in the region $0 < \hat{a} \leq 0.234$ (3s.f.).

Proof. The proof is identical to that of Theorem 5.1.1 with the only difference being the weightings of the components in the double sum, but this is simply derived by conditioning on the event, F, that there are no swaps between chains in the same region. □

Now return to the setting of a bi-modal target in two regions, given in Figure 5.4, and considering the same setup but instead deriving the optimal schedule via Corollary 5.2.2. Without the need for the plots it is clear that in this bimodal example, the optimal schedule is unaffected by the weight in mode 1, w_1, and instead is only focused on finding the optimal value of ℓ such that it maximises

$$2\ell^2 \Phi_{(0,1)} \left(-\frac{\ell \sigma_{1,2}(\beta)}{2} \right),$$

which is well established to induce a 0.234 rule for the acceptance of the relevant useful type swap moves and indeed the spacing will only depend on the value of $\sigma_{1,2}(\beta) = (I_1(\beta) + I_2(\beta))^{\frac{1}{2}}$. Consequently, for any value of $w_1 \in (0,1)$, the optimal spacing, $\hat{\ell}$ will remain the same and is bounded between the optimal spacings induced by $\sigma_{1,2}(\beta) = (2I_1(\beta))^{\frac{1}{2}}$ and $\sigma_{1,2}(\beta) = (2I_2(\beta))^{\frac{1}{2}}$. This can be considered as being a value for an optimal spacing with a compromise between the two extremes where the modes induce very different spacings for non useful swap moves.

In a bimodal example it is therefore very nice that the optimal spacings don't depend on the weights of the regions and only focuses on optimising the useful swap type moves independently of the regional weightings. However, this is clearly no longer the case when there are more than 2 regions, there is still the intuitively nice

property that if there was a dominant region then it's influence on the scaling won't be as powerful any longer.

The practicality of scaling spacings according to the result in Corollary 5.2.2 is limited but not impossible. In the colder states in the setting of well separated modes then local gradient information or a clustering approach could assign chain locations to modes but it is questionable whether this extra expense is worthwhile.

The major message from this section is that the $ESJD_\beta$ can be a misleading metric to optimise. It requires marginalisation of the temperature components by integrating out the dependency on the chain locations under the assumed infinitely fast mixing at each level. This is a highly unrealistic assumption and motivates a full joint location-temperature asymptotic analysis.

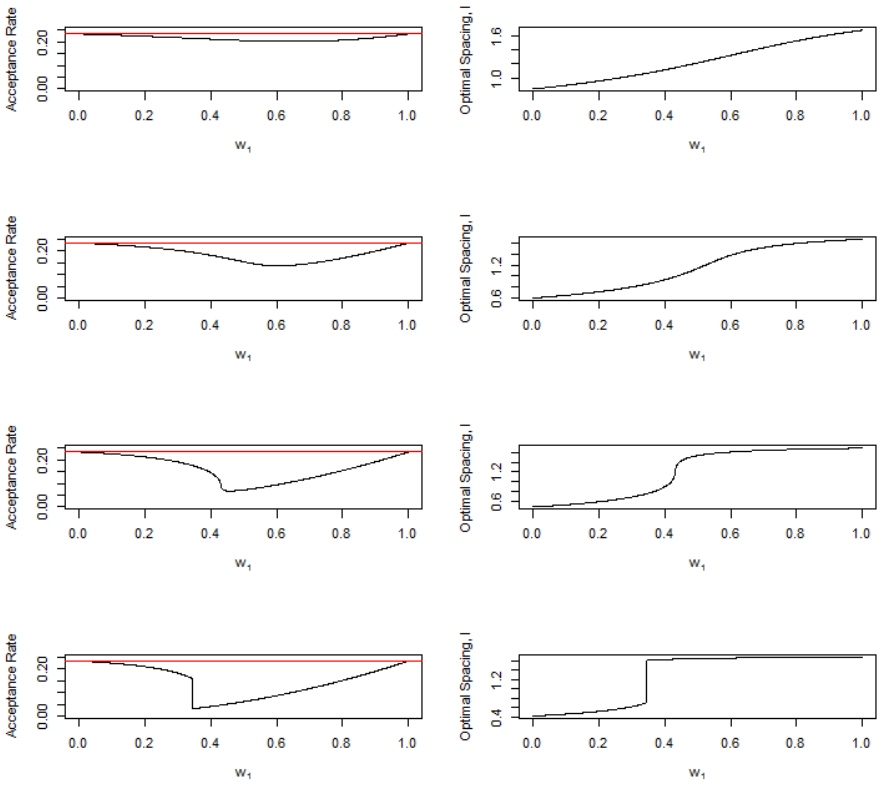

Figure 5.4: In the setting of Theorem 5.1.1, with 2 regions and for a fixed dimension assumed large, the optimal scalings and corresponding optimal acceptance rates are computed as the weight in the first region w_1 varies over the range 0 to 1. Furthermore within the respective regions $I_1(\beta) = 1$ and that $I_2(\beta) = V$ where a range of $V \in \{4, 8, 12, 16\}$ is considered. Each row of plots considers each value of V respectively. Left: Optimal acceptance rate plots. Right: the corresponding optimal spacings in terms of the scaling parameter l.

Chapter 6

Conclusions and Furtherwork

6.1 Conclusion

This thesis explored two core concepts relating to the performance of the PT algorithm. These were regarding improvements to the mixing speed of the hot state mixing information through the temperature schedule once a mode had been found (Chapter 2), and regarding the issues arising from the typical lack of regional weight preservation under power-tempering (Chapter 4). Each of these chapters gave rise to a prototype algorithm, QuanTA and HAT respectively, giving the first steps towards a more robust, scalable framework for tackling the issue of multi-modality in MCMC. Each of the new algorithms were accompanied by theoretical results that provide (asymptotic in dimension) optimal setup and these were established in Chapters 3 and 5 respectively.

Chapter 2 explored the restrictive nature of the spacings for consecutive levels of the temperature schedule. This limits the ambition of the proposal magnitude for jumps in the temperature space. It was shown that, for a unimodal setting, appropriate quantile preservation between levels allowed swap moves that were accepted with probability one. In general, a reparametrisation scheme preserving the quantile within a mode upon proposing a swap of the chain's value to a different temperature level is unfeasible. However, in the case of a Gaussian unimodal target there is a simple reparametrisation that can be carried out if one knows the location of the mode point. For this canonical example, it is possible to entirely overcome the curse of dimensionality that typically restricts the ambition of the inverse temperature spacings (which are usually $O(d^{-1/2})$) in a PT algorithm. In fact, in the Gaussian setting jumps in the temperature space of arbitrary distance are accepted with probability one.

It is this reparametrisation that motivated the design of a new prototype algorithm, QuanTA in Section 2.6, that attempts to utilise the Gaussian driven reparametrisation to approximately preserve the quantile within a mode upon proposing a swap to an adjacent temperature level. QuanTA exploits a population-based approach to MCMC that preserves the Markovian property whilst extracting information (in the form of mode points needed for the reparametrisation centrings) from the current values of a population of parallel Markov chains.

Empirical examples highlighted the vastly improved mixing that can be achieved by using the QuanTA approach in the canonical symmetric mode Gaussian mixture target setting, particularly as the dimensionality increased. Indeed, in this canonical setting it was illustrated that the vast number of intermediate temperatures that are typically required for the PT approach are unnecessary. Consequently, for cases where the local modes can be well approximated by a Gaussian distribution then the QuanTA approach can vastly improve the transfer speed of mixing information from the hot state to aid the cold state inter-modal mixing.

The QuanTA approach still has a number of key open problems; rendering it a prototype algorithm rather than a finished product. The QuanTA algorithm as specified in Section 2.6 requires prior specification of the number of modes, K, which for many practical problems is unrealistic. When the modes are critically not symmetric then the population size required for a stable robust method will likely grow exponentially with the dimension of the problem. The reparametrisation form is specific to a Gaussian mode and is inappropriate for modes that are not well approximated by a Gaussian distribution.

In practice the method may need added robustification by incorporating other reparametrisation shifts such as the shift required if the modal structures are Laplace; this gave rise to the development of QuanTAR in Section 2.10.3 which uses a random scan approach to using a reparametrisation move from a collection of pre-specified reparametrisations. Alas, these pre-specified forms will be limited, and this was illustrated in Section 2.10.2 for the t-distribution case when the ideal reparametrisation is only implicitly available.

Chapter 3 provided an accompanying result suggesting an approach for optimal setup of the QuanTA temperature schedule. The major result is given in Theorem 3.2.1 and this suggests that the optimal spacings in the inverse temperature schedule for QuanTA for a general target of iid form is still $O(d^{-1/2})$ (as is the suggested optimum for the PT approach); hence no generic improvement in the scalability with dimension. Importantly, Theorem 3.2.1 suggests that the temperature schedule setup can be tuned to optimality by aiming for a schedule that induces a

0.234 acceptance rate for swap move proposals between adjacent temperature levels.

Theorem 3.4.1 in Section 3.4 is supportive of the use of QuanTA; proving that it can exhibit higher order behaviour with respect to the spacings than the traditional PT setup does. The result describes the higher order spacings possible at super-cold temperatures in symmetric and asymmetric modal cases.

QuanTA only aims to aid mixing through the temperature schedule once a mode has been found. The work in Woodard *et al.* [2009a] and Woodard *et al.* [2009b], overviewed in Section 1.6, highlighted that the scalability of the PT approach using power-based tempering is poor for most interesting examples. Indeed, this is the case for the canonical setting of a Gaussian mixture model with modes that have differing covariance structures; overcoming this issue is the focus of Chapter 4. The major problem is the lack of regional weight preservation when using power-based tempering; a problem that is emphasised with an increase in dimensionality.

Using the Gaussian mixture setting as the canonical setting once more, the concept of an "ideal" tempered target was introduced in Section 4.2. Particularly through the colder temperature levels this approximately preserves the regional weight in a modal region. By approximating this ideal target, a prototype algorithm, HAT, was developed in Section 4.4 that essentially used rescaling regionally about the local mode point.

This approach appears to give impressive improvement in mixing speeds (per chain iteration) on difficult low to mid-dimensional examples presented. However, there are clear issues that render the current algorithm prototype rather than a finished product. Computationally, the algorithm is very expensive and requires the computation and inversion of hessian matrices at every evaluation of the target making the algorithm at least $O(d^3)$ in complexity.

However, it should be noted that recent collaborative work discussed in Section 4.5.1 suggests that in the Gaussian mixture setting then even with the hessian expense, the HAT algorithm is polynomially degrading with dimension as opposed to the exponential decay for the traditional PT approach. This work assumes immediate mixing at the hottest state but this is currently a major issue with the HAT approach, as is highlighted in Section 4.6. In high-dimensional settings, using RWM type moves will lead to exponentially slow mixing and likely mean that the capacitance condition in Woodard *et al.* [2009b], which is one aspect of a trio of conditions guaranteeing torpid mixing, would be satisfied. Chapter 4 has shown that weight preservation makes a huge difference to the robustness of the PT algorithm. Indeed, a major conclusion of Chapter 4 is that practitioners should be very wary

of using temperature swap move acceptance rates to diagnose the performance of the PT algorithm.

Chapter 5 provides an optimality result, Theorem 5.1.1, that accompanies the HAT algorithm (but is more broadly applicable to any regional power-based tempered weight preserving PT algorithm). Similar to Theorem 3.2.1 for QuanTA, Theorem 5.1.1 suggests an optimal setup to the temperature spacings for a regionally weight preserved PT algorithm for high dimensional settings. Again, the optimal schedule has temperature spacings that scale with dimension as $O(d^{-1/2})$. Importantly, Theorem 5.1.1 suggests that the temperature schedule setup can be tuned to optimality by aiming for a schedule that induces an acceptance rate in the range $0 < \hat{a} \leq 0.234$ for swap move proposals between adjacent temperature levels. Indeed, under the setting of a Gaussian mixture, Corollary 5.2.1 to Theorem 5.1.1, suggests that a geometric schedule is optimal; an extension to the geometric optimality for a Gaussian suggested in Atchadé *et al.* [2011].

Complementary to this, Section 5.2.1 discussed the appropriateness of using $ESJD_\beta$ as the measure of performance of the algorithm with regards to the success of inter-modal mixing. It showed that the use if the $ESJD_\beta$ is inappropriate if used in combination with the assumption of infinitely fast within temperature mixing. It produces an algorithm that perhaps can move effectively through the temperature schedule but can have critically bad performance with regards to inter-modal mixing.

To conclude, this thesis has made a contribution to improving arguably the most successful tool that practitioners use when implementing an MCMC algorithm in multi-modal settings. The prototype algorithms are accompanied with optimality theorems that are not only useful for algorithmic setup, but more crucially, provide insight into the scalability and merits of the approaches. However, the issue of multi-modality in MCMC is far from resolved; this thesis has highlighted core problems for future research and novel approaches that take the first steps to designing more robust scalable algorithms in this setting.

6.2 Further Work

Beyond the detailed shortcomings of the new algorithms that are presented in this thesis, two interesting directions for further work have become apparent. Both have clear motivation to overcoming the scalability issues with the PT and ST algorithms and are heavily linked to the QuanTA and HAT algorithms respectively.

6.2.1 Tempering with Implicit MCMC

Chapter 2 introduced QuanTA, utilising a deterministically reparametrised location temperature swap move to improve the mixing through the temperature schedule. For optimal and stable performance, QuanTA relies on knowing the number of modes prior to running the algorithm and this is typically unrealistic. This motivates a more robust approach. Section 2.5 eluded to using a Dirichlet process prior on the number of modes and performing a more robust clustering scheme than the basic K-means approach but this would inevitably be highly computationally intensive.

Intuitively, the Gaussian-driven reparametrisation draws the location in towards the mode point upon proposing a location that is colder, and for reversibility it repels the location away from the mode point when proposing a swap to a hotter temperature level; all with the hope of approximately preserving the quantile of the local mode. In a unimodal Gaussian setting, the direction towards the mode can be found with the use of local, log-target, gradient information to second order. This means that no population-based approach is required to estimate the mode point.

An approach that could be use to exploit this for the general multi-modal framework, whilst importantly still guaranteeing reversibility, is to use an implicit MCMC framework, e.g. Casella *et al.* [2011], which has been used to try to overcome some of the instability issues prevalent in the popular MALA approach when targeting light tailed distributions.

Consider the joint move setting of Chapter 2, which in QuanTA takes the form of a deterministic move for the location, x to a reparametrised version $x^{'}$:

$$(x, \beta) \rightarrow (x^{'}, \beta^{'}).$$

Suppose that this new location is given by solving the following implicit equation

$$x^{'} + g(\beta^{'}, x^{'}) \bigtriangledown \log \pi(x^{'}) = x + h(\beta, x) \bigtriangledown \log \pi(x) \tag{6.1}$$

where g and h are functions that must be specified. Solutions to this can be found using the Newton algorithm for finding the roots of such implicit equations. Assuming uniqueness of the solution, then reversibility is guaranteed.

From the work in Chapter 2 it seems natural that the choice of h and g should allow for swap moves between adjacent temperatures with arbitrary separation to be accepted with probability one for a uni-modal Gaussian target distribution. Hence,

setting

$$h(x,\beta) = g(x,\beta) = \left(\beta^{1/2} - 1\right)\left[-\bigtriangledown\bigtriangledown^{T}\log\pi(x)\right]^{-1} \tag{6.2}$$

then the solution to the implicit equation in the setting where the target at the cold state is a Gaussian with mean μ and variance matrix Σ becomes

$$\begin{aligned} x^{'} + g(\beta^{'}, x^{'}) \bigtriangledown \log\pi(x^{'}) &= x + h(\beta, x) \bigtriangledown \log\pi(x) \\ x^{'} + \left((\beta^{'})^{1/2} - 1\right)\Sigma\Sigma^{-1}(x^{'} - \mu) &= x + \left(\beta^{1/2} - 1\right)\Sigma\Sigma^{-1}(x - \mu) \\ x^{'} &= \left(\frac{\beta}{\beta'}\right)^{1/2}(x - \mu) + \mu. \end{aligned}$$

which mirrors the reparametrised location suggested for a Gaussian unimodal target in Chapter 2. For reversibility, it would be ideal that the solution is unique. To guarantee this, one would have to run the solver, e.g. the Newton scheme, from the proposed new location to ensure that one returns to the initial point x, doubling the computational overhead of the solution step.

A major issue that could prove problematic in this approach is the requirement for calculation of third derivatives. This is due to the deterministic nature of the move requiring a Jacobian for the transformation which is needed in the accept-reject step of the swap move. Defining,

$$b(x) := \left[-\bigtriangledown\bigtriangledown^{T}\log\pi(x)\right]^{-1} \bigtriangledown \log\pi(x),$$

then the Jacobian for the transformation $x \to x^{'}$ from solving the implicit equation (6.1) is given by

$$J_x(x^{'}) := \frac{\partial x^{'}}{\partial x} = \left[I_d + \left((\beta')^{1/2} - 1\right) J_{x'}(b(x^{'}))\right]^{-1}\left[I_d + \left((\beta')^{1/2} - 1\right) J_x(b(x))\right]$$

where I_d is the $d \times d$ identity matrix. If the problem doesn't have tractable gradient terms then this is certainly a non-trivial computational overhead. Early, attempts to run this on simple 1-dimensional Gaussian mixtures have shown similar inferential performance to the QuanTA approach but the scalability is poor with dimensionality. However, the concept of using gradient information to construct a joint temperature-location swap move seems to be an interesting future path for making QuanTA more robust.

6.2.2 Rapid Mixing via State Space Augmentation

The main message from the work in Chapter 4 which essentially follows from the work in Woodard *et al.* [2009b], is that when modes have differing scalings then naive tempering approaches scale badly with dimensionality. Woodard *et al.* [2009b] highlighted a property of persistence, see Section 1.6 for an overview. This characterizes the regional weight indifference between the cold and hot states when tempering. Chapter 4 demonstrates that this is a significant problem, even for low dimensional examples for Gaussian mixture target distributions. The HAT algorithm attempts to overcome the issue of regional weight preservation, but there are open issues with mixing at the hottest temperature levels, likely making the algorithm impractical for high dimensional problems.

Woodard *et al.* [2009a] showed that modal symmetry in a Gaussian mixture rendered the PT algorithm rapidly mixing. The HAT algorithm preserves weight but there can then be significant entropy changes between regions at the hottest level, inherent from the asymmetry of modal structure.

This motivates looking for a different approach that still aims to preserve regional weight but allows rapid mixing at the hottest temperature levels. Albeit at an early stage of development the thesis has pointed towards a promising approach that utilises the idea of modal symmetry and hence regional weight preservation upon tempering. Consider the following 1-dimensional bi-modal Gaussian mixture distribution with means at -10 and 10; variances of 0.1 and 9; and equal weighting, illustrated in Figure 6.1.

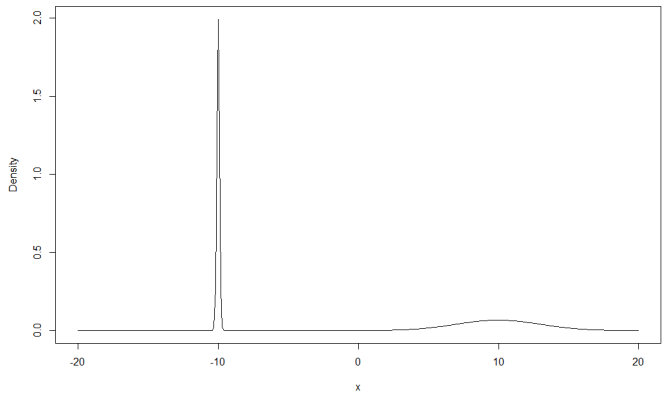

Figure 6.1: A 1-dimensional bi-modal Gaussian mixture distribution with means at -10 and 10; variances of 0.1 and 9; and equal weighting

Heuristically, it will always be an issue attempting to force a reversible Markov chain in the mode centred on -10 to the mode centred on 10 due to the dramatic change in entropy. However, consider augmenting the state space by a variable y which is designed in a way such that, in a Gaussian setting, the modes have an identical determinant of the covariance structure. Clearly, this will come in the form of a conditional distribution given the current value of the "x" variable and could crudely be described through the gradient information from the point x, similar to the HAT approach.

In the case of the Gaussian mixture in Figure 6.1 then the ideal is that the new target would (approximately) be the bi-modal bi-variate Gaussian target:

$$0.5 \times N\left(\begin{bmatrix} -10 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}\right) + 0.5 \times N\left(\begin{bmatrix} 10 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}\right) \tag{6.3}$$

which has a density illustrated in Figure 6.3.

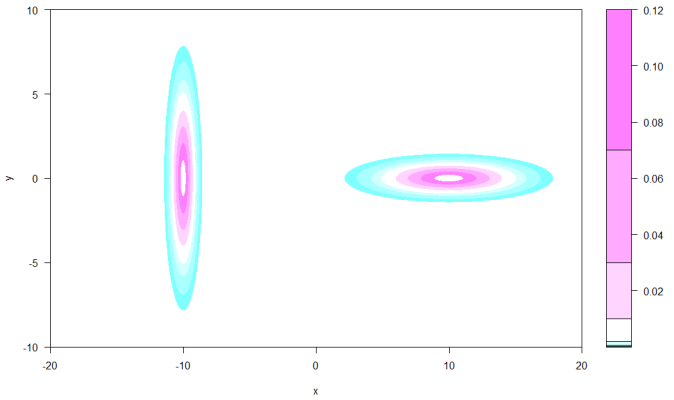

Figure 6.2: Density contour plot for the bi-modal bi-variate Gaussian target given in equation (6.3).

Hence the core idea is that through state space augmentation, the hope is that the mode points evaluations can be on the scale of the weight of the mode rather than depending on the covariance structure. Hence, upon tempering there shouldn't be an exponential decay of regional weight indifference.

Bibliography

Al-Awadhi, F., Hurn, M. and Jennison, C. (2004) Improving the Acceptance Rate of Reversible Jump MCMC Proposals. *Statistics & Probability letters*, **69**, 189–198.

Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2007) Convergence of the Equi-energy Sampler. In *ESAIM: Proceedings*, vol. 19, 1–5. EDP Sciences.

Atchadé, Y. F. and Liu, J. S. (2004) The Wang-Landau Algorithm for Monte Carlo Computation in General State Spaces. *Statistica Sinica*, **20**, 209–33.

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**, 555–568.

Barndorff-Nielsen, O. E. and Nielsen, O. E. B. (1989) Asymptotic Techniques; for use in Statistics. Tech. rep.

Behrens, G. R. (2008) *Mode Jumping in MCMC*. Ph.D. thesis, University of Bath.

Bottou, L. (2010) Large-scale Machine learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, 177–186. Springer.

Bradley, P. S. and Fayyad, U. M. (1998) Refining Initial Points for K-Means Clustering. *ICML*, **98**, 91–99.

Brooks, S. P., Giudici, P. and Roberts, G. O. (2003) Efficient Construction of Reversible Jump Markov chain Monte Carlo Proposal Distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 3–39.

Casella, B., Roberts, G. and Stramer, O. (2011) Stability of partially implicit langevin schemes and their mcmc variants. *Methodology and Computing in Applied Probability*, **13**, 835–854.

Durrett, R. (2010) *Probability: Theory and Examples*. Cambridge university press.

Elías, C. and Del Campob, P. C. (2007) Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. *Revista colombiana de estadística*, **30**, 231–245.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The Elements of Statistical Learning*, vol. 1. Springer Series in Statistics New York.

Geyer, C. J. (1991) Markov chain Monte Carlo Maximum Likelihood. *Computing Science and Statistics*, **23**, 156–163.

Geyer, C. J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, 473–483.

Gilbert, P., Gilbert, M. P. and Varadhan, R. (2006) The numderiv package.

Gilks, W. R., Roberts, G. O. and George, E. I. (1994) Adaptive Direction Sampling. *The Statistician*, 179–189.

Girolami, M. and Calderhead, B. (2011) Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 123–214.

Gramacy, R., Samworth, R. and King, R. (2010) Importance Tempering. *Statistics and Computing*, **20**, 1–7.

Green, P. J. and Mira, A. (2001) Delayed Rejection in Reversible Jump Metropolis–Hastings. *Biometrika*, **88**, 1035–1053.

Grimmett, G. and Stirzaker, D. (2001) *Probability and random processes*. Oxford university press.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A k-means Custering Algorithm. *Applied statistics*, 100–108.

Hastie, D. (2005) *Towards Automatic Reversible Jump Markov chain Monte Carlo*. Ph.D. thesis, University of Bristol.

Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markov chains and their Applications. *Biometrika*, **57**, 97–109.

Jasra, A., Stephens, D. A. and Holmes, C. C. (2007) Population-based Reversible Jump Markov chain Monte Carlo. *Biometrika*, **94**, 787–807.

Jennison, C. and Sharp, R. (2006) Mode Jumping in MCMC: Adapting Proposals to the Local Environment.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. (2002) An Efficient k-means clustering algorithm: Analysis and Implementation. *IEEE transactions on pattern analysis and machine intelligence*, **24**, 881–892.

Kim, S., Tadesse, M. G. and Vannucci, M. (2006) Variable Selection in Clustering via Dirichlet process Mixture Models. *Biometrika*, **93**, 877–893.

Kipnis, C. and Varadhan, S. S. (1986) Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, **104**, 1–19.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. *et al.* (1983) Optimization by Simulated Annealing. *science*, **220**, 671–680.

Kone, A. and Kofke, D. A. (2005) Selection of Temperature Intervals for Parallel-Tempering Simulations. *The Journal of Chemical Physics*, **122**, 206101.

Kou, S., Zhou, Q. and Wong, W. H. (2006) Equi-energy Sampler with Applications in Statistical Inference and Statistical Mechanics. *The Annals of Statistics*, 1581–1619.

Livingstone, S. (2015) Geometric Ergodicity of the Random Walk Metropolis with Position-Dependent Proposal Covariance. *arXiv preprint arXiv:1507.05780*.

Madras, N. and Zheng, Z. (2003) On the Swapping Algorithm. *Random Structures & Algorithms*, **22**, 66–97.

Marinari, E. and Parisi, G. (1992) Simulated Tempering: a New Monte Carlo Scheme. *EPL (Europhysics Letters)*, **19**, 451.

Meng, X.-L. and Schilling, S. (2002) Warp Bridge Sampling. *Journal of Computational and Graphical Statistics*, **11**, 552–586.

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**, 1087–1092.

Meyn, S. P. and Tweedie, R. L. (2012) *Markov Chains and Stochastic Stability*. Springer Science & Business Media.

Miasojedow, B., Moulines, E. and Vihola, M. (2013) An Adaptive Parallel Tempering Algorithm. *Journal of Computational and Graphical Statistics*, **22**, 649–664.

Neal, R. M. (1996) Sampling from Multimodal Distributions using Tempered Transitions. *Statistics and Computing*, **6**, 353–366.

Neal, R. M. (2000) Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Neal, R. M. (2001) Annealed Importance Sampling. *Statistics and Computing*, **11**, 125–139.

Nemeth, C., Lindsten, F., Filippone, M. and Hensman, J. (2017) Pseudo-extended Markov Chain Monte Carlo. *ArXiv e-prints*.

Øksendal, B. (2003) Stochastic Differential Equations. In *Stochastic Differential Equations*, 65–84. Springer.

Olver, F. (1968) Error bounds for the Laplace Approximation for Definite Integrals. *Journal of Approximation Theory*, **1**, 293–313.

Papaspiliopoulos, O. and Roberts, G. O. (2003) Non-centered Parameterisations for Hierarchical Models and Data Augmentation. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, 307–326. Oxford University Press, USA.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 59–73.

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**, 4119–4128.

Raykov, Y. P., Boukouvalas, A., Little, M. A. *et al.* (2016) Simple approximate MAP inference for Dirichlet processes Mixtures. *Electronic Journal of Statistics*, **10**, 3548–3578.

Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 400–407.

Roberts, G. and Gilks, W. (1994) Convergence of Adaptive Direction Sampling. *Journal of Multivariate Analysis*, **49**, 287–298.

Roberts, G. O., Gelman, A., Gilks, W. R. *et al.* (1997) Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability*, **7**, 110–120.

Roberts, G. O. and Rosenthal, J. S. (2007) Coupling and Ergodicity of Adaptive Markov chain Monte Carlo Algorithms. *Journal of Applied Probability*, 458–475.

Roberts, G. O. and Rosenthal, J. S. (2009) Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349–367.

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**, 131–149.

Roberts, G. O., Rosenthal, J. S. *et al.* (2001) Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, **16**, 351–367.

Roberts, G. O., Rosenthal, J. S. *et al.* (2004) General State Space Markov Chains and MCMC Algorithms. *Probability Surveys*, **1**, 20–71.

Roberts, G. O. and Sahu, S. K. (1997) Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 291–317.

Roberts, G. O. and Stramer, O. (2002) Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology and Computing in Applied Probability*, **4**, 337–357.

Schreck, A., Fort, G. and Moulines, E. (2013) Adaptive equi-energy sampler: Convergence and Illustration. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **23**, 5.

Sherlock, C. (2006) *Methodology for Inference on the Markov Modulated Poisson Process and Theory for Optimal Scaling of the Random Walk Metropolis*. Ph.D. thesis, Lancaster University.

Tak, H., Meng, X.-L. and van Dyk, D. A. (2016) A Repulsive-Attractive Metropolis Algorithm for Multimodality. *arXiv preprint arXiv:1601.05633*.

Tjelmeland, H. and Hegstad, B. K. (2001) Mode Jumping Proposals in MCMC. *Scandinavian Journal of Statistics*, **28**, 205–223.

VanDerwerken, D. N. and Schmidler, S. C. (2013) Parallel Markov Chain Monte Carlo. *arXiv preprint arXiv:1312.7479*.

Wang, F. and Landau, D. (2001) Determining the Density of States for Classical Statistical Models: A Random Walk Algorithm to Produce a flat Histogram. *Physical Review E*, **64**, 056101.

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009a) Conditions for Rapid Mixing of Parallel and Simulated Tempering on Multimodal Distributions. *The Annals of Applied Probability*, 617–640.

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009b) Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering. *Electronic Journal of Probability*, **14**, 780–804.

Žalik, K. R. (2008) An Efficient K-means Clustering Algorithm. *Pattern Recognition Letters*, **29**, 1385–1391.

Zhao, W., Ma, H. and He, Q. (2009) Parallel k-means Clustering Based on Mapreduce. In *IEEE International Conference on Cloud Computing*, 674–679. Springer.

Zheng, Z. (2003) On Swapping and Simulated Tempering Algorithms. *Stochastic Processes and their Applications*, **104**, 131–154.