Applied Probability Trust (April 30, 2019)

Accepted for publication by the *Applied Probability Trust* (http://www.appliedprobability.org) in *Advances in Applied Probability* 51.3 (September 2019).

ACCELERATING PARALLEL TEMPERING: QUANTILE TEMPERING ALGORITHM (QUANTA)

NICHOLAS G. TAWN,* *University of Warwick*

GARETH O. ROBERTS,** *University of Warwick*

Abstract

It is well-known that traditional MCMC methods can fail to effectively explore the state space for multimodal problems. Parallel tempering is a well-established population approach for such target distributions involving a collection of particles indexed by temperature. However this method can suffer dramatically from the curse of dimensionality. This paper introduces an improvement on parallel tempering called QuanTA. A comprehensive theoretical analysis quantifying the improved efficiency and scalability of the approach is given. Under weak regularity conditions, QuanTA gives accelerated mixing through the temperature space. Empirical evidence of the effectiveness of this new algorithm is illustrated on canonical examples.

Keywords: Parallel Tempering; Metropolis coupled MCMC; Accelerated MCMC; Markov Chain Monte Carlo; Multimodal; Population-MCMC

2010 Mathematics Subject Classification: Primary 60J22

Secondary 65C05;65C60

1. Introduction

Consider the problem of stochastic simulation from a target distribution, $\pi(x)$ on a d-dimensional state space $\mathcal{X}$ where $\pi(\cdot)$ is known up to a scaling constant. The *gold standard* methodology for this problem uses Markov chain Monte Carlo (MCMC).

* Postal address: Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

** Postal address: Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

However these methods often perform poorly in the context of multimodality.

Most MCMC algorithms use localised proposal mechanisms, tuned towards local approximate optimality e.g., [31], [35]. Indeed many MCMC algorithms incorporate local gradient information in the proposal mechanisms, typically attracting the chain back towards the centre of the mode. This can exacerbate the difficulties of moving between modes, [18].

Popular methods used to overcome these issues include *simulated tempering*, [19] and the population-based version, *parallel tempering*, [10], [11]. These methods use state space augmentation to allow Markov chains to explore target distributions proportional to $\pi^{\beta}(x)$ for β typically in the range $(0, 1]$. For simulated tempering this is done by introducing an auxiliary inverse temperature variable, β, and running a $(d+1)$-dimensional Markov chain on $\mathcal{X} \times \Delta$, where Δ consists of a discrete collection of possible inverse temperatures including 1. For the more practically applicable parallel tempering approach, a Markov chain is run on a $(|\Delta| \times d)$-dimensional state space, $\mathcal{X}^{|\Delta|}$, where $|\Delta|$ denotes the cardinality of the set Δ.

Within this paper we will concentrate on parallel tempering as it obviates the need to approximate certain normalisation constants to work effectively. While parallel tempering has been highly successful, for example see [21], [49], [5] etc, its efficiency declines as a function of d, at least linearly and often much worse [1] and [48]. This is caused by the need to set inter-inverse temperature spacings in Δ extremely small to make swaps between temperatures feasible.

This paper will introduce and analyse the QuanTA algorithm which facilitates inter-temperature swaps by proposing moves which attempt to adjust within-mode variation appropriately for the proposed new temperature. This leads to improved temperature mixing, which in turn leads to vastly improved inter-modal mixing. Its typical improvement is demonstrated in Figure 1 with a 5-mode target distribution. The construction of QuanTA resonates with the non-centering MCMC methodology described, for example in [26], [4], [14] and [27].

Supporting theory is developed to guide setup and analyse the utility of the novel QuanTA scheme. There are two key theoretical results. The first, Theorem 1, establishes that there is an optimal temperature schedule setup for QuanTA; concluding that in general the dimensionality scaling of the distance between consecutive inverse

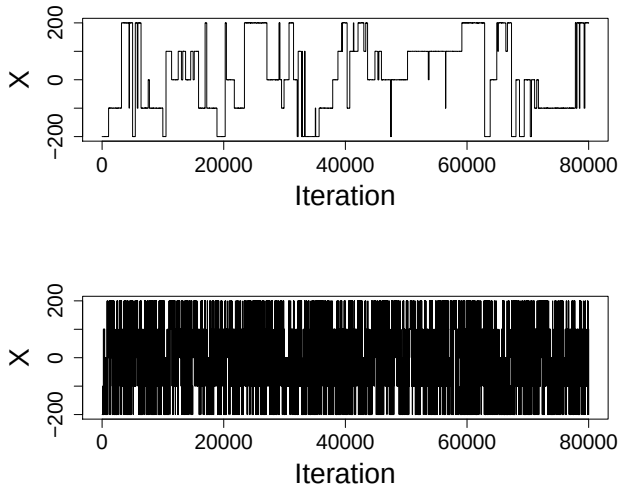

FIGURE 1: Trace plots of the target state chains for representative runs of the Parallel Tempering (top) and QuanTA schemes (bottom).

temperature spacings should be $\mathcal{O}(d^{-1/2})$. Further to this it suggests that optimising the expected squared jumping distance between any two consecutive temperature levels induces a temperature swap move acceptance rate of 0.234; giving a useful metric for a practitioner to optimally tune QuanTA. The second key theoretical contribution, Theorem 2, of this paper shows that, under mild regularity conditions, the optimal temperature spacings of QuanTA are more ambitiously spaced than for the standard parallel tempering algorithm for cold (i.e. large) values of the inverse temperatures. The significance of this result is that QuanTA can give accelerated mixing through the cooler parts of the temperature schedule by allowing more ambitious temperature spacings.

This paper is structured into 6 core sections. Sections 2 reviews the parallel tempering algorithm and some of the relevant existing literature. Section 3 motivates the main idea behind the novel QuanTA scheme, which is then presented in Section 4. QuanTA utilises a population MCMC approach that requires a clustering scheme; discussion for this is found in Section 5. Section 6 contains the core theoretical contributions mentioned above. Simulation studies are detailed in Section 7 along with a discussion of the computational complexity of QuanTA.

2. The Parallel Tempering (PT) algorithm

There is an array of methodology available to overcome the issues of multimodality using MCMC or closely related Sequential Monte Carlo (SMC) techniques, e.g. [46], [10], [19], [22], [17], [24] [28], [37]. The majority of these methods use state space augmentation approaches. Auxiliary distributions that allow a Markov chain to explore the entirety of the state space are targeted and their mixing information is then passed on to aid inter-modal mixing in the desired target. A convenient approach for the augmentation methods is to use power-tempered target distributions i.e., the target distribution at inverse temperature level, β, for $\beta \in (0, 1]$ is defined as

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

Such targets are the most common choice of auxiliary target when augmenting the state space for use in the popular simulated tempering (ST) and parallel tempering (PT) algorithms introduced in [19] and [10]. For each algorithm one needs to choose a sequence of $n+1$ "inverse temperatures", $\Delta = \{\beta_0, \ldots, \beta_n\}$, where $0 \leq \beta_n < \beta_{n-1} < \ldots < \beta_1 < \beta_0 = 1$ with the specification that a Markov chain sampling from the target distribution $\pi_{\beta_n}(x)$ can mix well across the entire state space.

The PT algorithm runs a Markov chain on the augmented state space, $\mathcal{X}^{(n+1)}$, targeting an invariant distribution given by

$$\pi_n(x_0, x_1, \ldots, x_n) \propto \pi_{\beta_0}(x_0)\pi_{\beta_1}(x_1)\ldots\pi_{\beta_n}(x_n). \tag{1}$$

From an initialisation point for the chain the PT algorithm alternates between two types of Markovian move. *Within temperature* Markov chain moves that use standard localised MCMC schemes to update each of the x_i whilst preserving marginal invariance. *Temperature swap* moves that propose to swap the chain locations between a pair of adjacent temperature components. It is these swap moves that will allow mixing information from the hot, rapidly-mixing temperature level to be passed to aid mixing at the cold target state.

To perform the swap move a pair of temperatures is chosen uniformly from the set of all adjacent pairs, call this pair x_i and x_{i+1} at inverse temperatures β_{x_i} and $\beta_{x_{i+1}}$ respectively. The proposal is then

$$(x_0, \ldots, x_i, x_{i+1}, \ldots, x_n) \to (x_0, \ldots, x_{i+1}, x_i, \ldots, x_n) \tag{2}$$

To preserve detailed balance and therefore invariance to $\pi_n(\cdot)$, the swap move is accepted with probability

$$A = \min\left(1, \frac{\pi_{\beta_{x_{i+1}}}(x_i)\pi_{\beta_{x_i}}(x_{i+1})}{\pi_{\beta_{x_i}}(x_i)\pi_{\beta_{x_{i+1}}}(x_{i+1})}\right). \tag{3}$$

It is the combination of the suitably specified *within temperature* moves and *temperature swap* moves that ensures ergodicity of the Markov chain to the target distribution, $\pi_n(\cdot)$. Note that the *within temperature* moves certainly influence the performance of the algorithm, [9]; however the focus of the work in this article will be on designing a novel approach for the *temperature swap* move.

The novel work presented in this paper focuses on the setting where the d-dimensional state space is given by $\mathbb{R}^d$ and the target, $\pi(\cdot)$, is the associated probability density function. Thus, herein take $\mathcal{X} = \mathbb{R}^d$ but note that natural generalisations to other state spaces and settings are possible.

3. Modal rescaling transformation

3.1. A motivating transformation move

Consider a PT algorithm that has two components x_1 and x_2 running at the neighbouring inverse temperature levels β and $\beta^{'}$. Suppose that a temperature swap move is proposed between the two chains at the two temperature levels. Due to the dependence between the location in the state space and the temperature level, β and $\beta^{'}$ need to be close to each other to avoid the move having negligible acceptance probability. Intuitively, the problem is that the proposal from the hotter chain is likely to be an "unrepresentative" location at the colder temperature and vice versa.

So there is clearly a significant dependence between the temperature value and the location of the chain in the state space; thus explaining why temperature swap moves between arbitrarily-largely spaced temperatures are generally rejected. This issue is typically exacerbated when the dimensionality grows.

Consider for motivational purposes, a simple one-dimensional setting where the state space is given by $\mathbb{R}$ and the target density is given by $\pi(\cdot)$. For notational convenience letting $j = i + 1$, suppose that a temperature swap move has been proposed between adjacent levels β_i and β_j with marginal component values x_i and x_j respectively.

Suppose an oracle has provided a function, $t_{ij} : \mathbb{R} \to \mathbb{R}$, that is bijective, with $t_{ji}(t_{ij}(x)) = x$, and differentiable and preserves the *cumulative distribution function*, CDF, between the two temperature levels such that

$$F_{\beta_j}(t_{ij}(x)) = F_{\beta_i}(x) \tag{4}$$

where $F_\beta(\cdot)$ denotes the CDF of $\pi^\beta(\cdot)$.

So suppose that rather than the standard temperature swap move proposal in (2), the following is instead proposed:

$$(x_0, \ldots, x_i, x_j, \ldots, x_n) \to (x_0, \ldots, t_{ji}(x_j), t_{ij}(x_i), \ldots, x_n) \tag{5}$$

To preserve detailed balance this is accepted with an acceptance ratio similar to reversible-jump MCMC, [12], to account for the deterministic transformation:

$$\min\left(1, \frac{\pi^{\beta_j}(t_{ij}(x_i))\pi^{\beta_i}(t_{ji}(x_j))}{\pi^{\beta_i}(x_i)\pi^{\beta_j}(x_j)} \left|\frac{\partial t_{ij}(x_i)}{\partial x}\right| \left|\frac{\partial t_{ji}(x_j)}{\partial x}\right|\right). \tag{6}$$

By differentiating (4) wrt x and rearranging

$$\frac{\pi_{\beta_j}(t_{ij}(x))}{\pi_{\beta_i}(x)} \left|\frac{\partial t_{ij}(x)}{\partial x}\right| = 1. \tag{7}$$

Using the result in (7), and noting the cancellation of normalising constants in (6), one can see that (6) evaluates to one and hence such a swap would always be accepted. Essentially, the transformation $t_{ij}(\cdot)$ has made the acceptance probability of a temperature swap move independent of the locations of x_i and x_j in the state space.

In practice, a CDF-preserving transformation $t_{ij}(\cdot)$ will not generally be available. Consider a simplified setting when the target is now a d-dimensional Gaussian, i.e. $\pi \sim N(\mu, \Sigma)$, and so the tempered target at inverse temperature β is given by $\pi^\beta \sim N(\mu, \Sigma/\beta)$. Defining a d-dimensional transformation by

$$t_{ij}(x, \mu) = \left(\frac{\beta_i}{\beta_j}\right)^{1/2} (x - \mu) + \mu, \tag{8}$$

a simple calculation shows that in this setting such a transformation, which only requires knowledge of the mode location, permits swap moves to always be accepted independently of the dimensionality and magnitude of the inverse temperature spacings.

In a broad class of applications it is not unreasonable to make a Gaussian approximation to posterior modes, [36]. Indeed this is the motivation for the similar transformation derived in [14] for use in a reversible-jump MCMC framework.

Beyond a single dimensional setting a quantile is typically non-unique since multiple locations in the state space produce the same CDF evaluation. However, the transformation in (8), motivated via Gaussianity, is bijective and can be seen as a way of uniquely pairing two quantiles at different temperature levels allowing for "quantile uniqueness" and therefore reversibility of the transformation. It is from this that QuanTA gets its name.

3.2. Transformation-aided move in a PT framework

In a multimodal setting a single Gaussian approximation to the posterior will be poor. However, it is often reasonable that the local modes may be individually approximated as Gaussian. This paper explores the use of the transformation in (8) applied to the local mode with the aim being to accelerate the mixing through the temperature schedule of a PT algorithm.

Now that the transformations are localised to modes one needs careful specification of the transformation function. Suppose that there is a collection of K mode points, $\mu_1, \ldots, \mu_K$ and a metric, $m(x, y)$ for $x, y \in \mathbb{R}^d$, that will be used to associate locations in the state space with a mode. To this end define the mode allocating function

$$Z(x) = \underset{h \in \{1,\ldots,K\}}{\arg\min} \left[m(x, \mu_h) \right].$$

Then, with $t_{ij}(\cdot)$ from (8) define the sets

$$A_{ij} = \left\{ x \in \mathbb{R}^d : Z(t_{ij}(x, \mu_{Z(x)})) = Z(x) \right\} \tag{9}$$

and define the transformation,

$$t(x, \beta_i, \beta_j) = t_{ij}(x, \mu_{Z(x)}). \tag{10}$$

Fundamentally, the set, A_{ij}, is the subset of the β_i-level marginal state space where the transformation in (10) is reversible, i.e., one would remain associated with the same mode point after the transformation. An illustrative example is given in Figure 2.

The aim is to use the transformation in a PT framework. So suppose that a temperature swap move proposal is made between two marginal components x_i and x_j

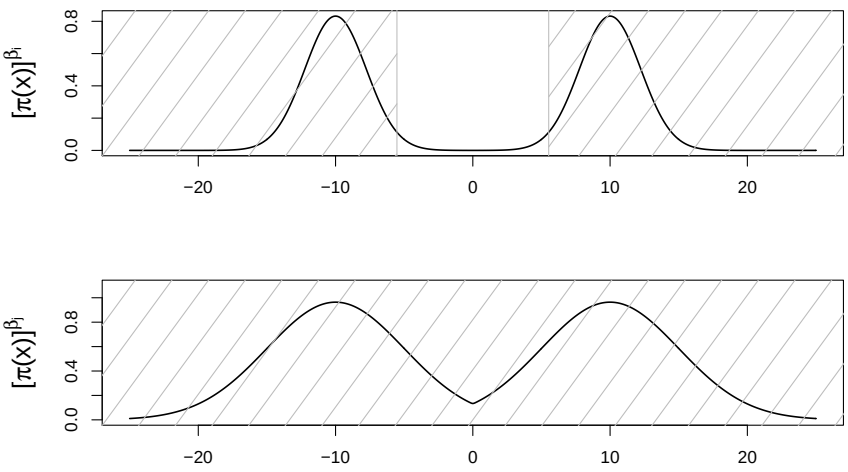

FIGURE 2: Plots of a simple bimodal Gaussian mixture distribution at two temperature levels β_i and β_j. The metric, $m(\cdot,\cdot)$, that allocates locations to mode points is chosen to be the Euclidean metric on $\mathbb{R}$. The shaded regions represent A_{ij} and A_{ji} in the top and bottom plots respectively.

at respective inverse temperatures β_i and β_j with $\beta_i > \beta_j$. The idea is that this swap move now utilises (10) so that the proposed move takes the form

$$(x_0, \ldots, x_i, x_j, \ldots, x_n) \to (x_0, \ldots, t(x_j, \beta_j, \beta_i), t(x_i, \beta_i, \beta_j), \ldots, x_n) \tag{11}$$

which to satisfy detailed balance is accepted with probability

$$\min\left(1, \frac{\pi(t(x_i, \beta_i, \beta_j))^{\beta_j} \pi(t(x_j, \beta_j, \beta_i))^{\beta_i}}{\pi(x_i)^{\beta_i} \pi(x_j)^{\beta_j}} \mathbb{1}_{\{x_i \in A_{ij}\}} \mathbb{1}_{\{x_j \in A_{ji}\}}\right). \tag{12}$$

Proposition 1. *Consider a Markov chain that is in stationarity with a target distribution given by* (1) *on a state space* $\mathcal{X} = \mathbb{R}^{(n+1)d}$. *Let* $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$. *If a temperature swap move of the form* (11) *is proposed where the transformation is given by* (10) *and is accepted with probability given in* (12) *then the chain is invariant with respect to* (1).

4. Quantile Tempering Algorithm (QuanTA)

The aim is is to utilise the transformation-aided temperature swap moves in a PT algorithm. However, to perform the transformation one needs a collection of K centring

points. Typically these are unknown a priori and must be estimated. There are three obvious directions: prior optimisation search for modes and then fix these points for the duration of the algorithm run; adaptively learn the mode points during the run of the algorithm; or use a population Monte Carlo approach where centres are estimated each time they are needed and using the most recent location of the Markov chain (thus preserving the Markov property) and then used in a way that ensures that the Markov chain remains reversible.

Fixing the centring points prior to the run of the algorithm can be highly non-robust since it leaves no scope to adjust in situations when new regions of mass are found during the run of the Markov chain. Also it requires careful and potentially expensive pre-computation.

Using adaptive techniques that "learn" the best centring points are therefore preferable. However, adaptive MCMC approaches where the proposal mechanism adapts throughout the run of the algorithm, [32], [33], requires careful design to ensure ergodicity. In contrast the population MCMC approach essentially allows the proposal mechanism to adapt at every iteration without concerns about affecting the ergodicity, albeit typically at a computational cost which can be somewhat mitigated by exploiting computer parallelism.

The algorithm presented below utilises a population MCMC approach with N "copies" of the PT approach to provide a large population of points in the state space that can be used in an appropriate clustering procedure to obtain a collection of K centring points; suggestions of which are given in Section 5.

4.1. The Procedure

QuanTA runs the equivalent of N parallel tempering algorithms procedures in parallel with each single procedure using the same tempering schedule. With a temperature schedule given by $\Delta = \{\beta_0, \ldots, \beta_n\}$, the QuanTA approach can be seen as running a single Markov Chain on the augmented state space, $(\mathbb{R}^d)^{n*N}$. Denoting $\mathbf{x} = (x_{(1,0)}, \ldots, x_{(1,n)}, x_{(2,0)}, \ldots, x_{(N,n)})$, the invariant target distribution for the Markov chain induced by QuanTA is

$$\pi_Q(\mathbf{x}) \propto \prod_{i=1}^{N} \pi_{\beta_0}\left(x_{(i,0)}\right) \pi_{\beta_1}\left(x_{(i,1)}\right) \ldots \pi_{\beta_n}\left(x_{(i,n)}\right).$$

Initialisation: to initialise the QuanTA algorithm, one is required to choose: initial starting values for the Markov chain components; a suitable temperature schedule (see Theorem 1 in Section 6.2 for suggested optimality criteria for the temperature schedule); the size of N and suitable parameters for the chosen clustering method that will be used.

Running the chain: from the start point of the chain, QuanTA alternates between two types of Markov chain moves.

Within temperature Markov chain moves that use standard localised MCMC schemes for marginal updates of each of the $x_{(i,j)}$. Essentially, this is just Metropolis-within-Gibbs MCMC and in this setting, with hugely exploitable marginal independence, this process is highly parallelisable. Denote the π_Q-invariant Markov transition kernel that performs temperature marginal updates on all components from a current point $\mathbf{x}$ as $P_1(\mathbf{x}, d\mathbf{y})$.

Temperature swap moves that propose to swap the chain locations between a pair of adjacent temperature components. This is where QuanTA differs from the standard PT procedure and uses the new transformation aided temperature swap move detailed in Section 3.2 in particular in (10). This follows a two phase population-MCMC update procedure.

- **Phase 1:** Group marginal components into two collections,

$$\begin{aligned} C_1 &= \{x_{(i,j)} : i = 1, \ldots, \lfloor N/2 \rfloor \text{ and } j = 0, \ldots, n\} \\ C_2 &= \{x_{(i,j)} : i = (\lfloor N/2 \rfloor + 1), \ldots, N \text{ and } j = 0, \ldots, n\}. \end{aligned}$$

  An appropriate clustering scheme (see Section 5) is performed on C_1 providing a set of K centres $\{c_1, \ldots, c_K\}$. To enhance the effectiveness of the transformation it is suggested that these cluster centre points are used as initialisation locations for a suitable local optimisation procedure to find K mode points $M_1 = \{\mu_1, \ldots, \mu_K\}$ of $\pi(\cdot)$ (see Theorem 2 in Section 6.2).

  For each $i \in \{(\lfloor N/2 \rfloor + 1), \ldots, N\}$, sample $l \sim Unif\{0, 1, \ldots, n-1\}$ and select the corresponding pair of adjacent temperature marginals $(x_{(i,l)}, x_{(i,l+1)})$ for a temperature swap move proposal utilising the transformation from (10) (which is centred on the associated point from M_1). This move is accepted with probability (12).

- **Phase 2:** Repeat phase 1 but with the roles of C_1 and C_2 reversed.

Denote the π_Q-invariant Markov transition kernel that implements this temperature swap update procedure for all components using the above two-phase process by $P_2(\mathbf{x}, d\mathbf{y})$.

From the initialisation point $\mathbf{x}$ then the Markov chain output is created by application of the following kernel compilation:

$$(P_2 \circ P_1^k)^T$$

where k is the user-chosen number of within temperature Markov chain updates between each swap move proposal and T is the user-chosen number of iterations of the algorithm before stopping.

Remark 1. By Proposition 1 the transformation-aided temperature swap move established in Section 3.2 leaves the Markov chain constructed by QuanTA $\pi_Q(\cdot)$ invariant. Provided appropriate within-temperature MCMC moves are utilised then QuanTA establishes an irreducible and aperiodic $\pi_Q(\cdot)$ invariant Markov chain as required, [42].

Remark 2. Although not made explicit in the notation above, the transformation centring points, $\{\mu_1, \ldots, \mu_K\}$, are repeatedly estimated from subsets of the points in the population of the N replicate PT schemes. As such their values are not fixed and have flexibility to change and adapt accordingly with the population. This is a major bonus of the population MCMC framework for a multimodal problem; allowing powerful and robust adaptation when new regions of mass are discovered.

5. Estimating local mode locations

The QuanTA algorithm, presented in Section 4, requires online estimation of the local mode points as centring location for the transformation. This section outlines a practical scheme that is used in the canonical simulation studies.

With a typically unknown number of modes and a population of chains, a principled approach would be to fit a Dirichlet Process mixture model, e.g. [23] and [16]. A comprehensive Gibbs sampling approach for this can be computationally expensive, but there are alternative cheaper but approximate methods that are left for exploration in further work, [29].

For the examples with well-separated modes that were studied here it sufficed to use a cheap and fast clustering scheme, [8]. To this end a K means approach was used, [13], where although the theoretical computational complexity is NP-hard in full generality, when in a setting with well-separated clusters the algorithm rapidly reaches a good solution as the problem is much easier.

The clustering procedure provides a collection of cluster centres that can be directly used as centring points for the transformation or as very useful initialisation points for a local optimisation method. Indeed, Theorem 2 of Section 6 shows that QuanTA can achieve accelerated mixing through the temperature levels when the centring point is chosen as the mode point, particularly at colder temperatures when the Gaussian approximation to the mode becomes increasingly accurate, e.g. [2] and [25].

5.1. A weighted K means clustering

Typically the K means algorithm assigns all points equal leverage in determining cluster centres. A weighted K means approach incorporates weights that can alter the leverages of points. In the tempering setting chains at the colder states, where the modes are less disperse, should have more leverage in determining the centres.

Weighted K means is an almost identical procedure to the K means algorithm of [13] but now incorporates the weights to give points leverage. For the setting of interest each chain location will be allocated a weight, determined by their inverse temperature value. So for a collection of ν chain locations $x_1, \ldots, x_\nu$ at inverse temperature levels $\beta_{x_1}, \ldots, \beta_{x_\nu}$ form their respective weights.

The weighted K means algorithm gives a pragmatic way of creating a plausible particle allocation, $\hat{S}$, in to K partitions. Denoting a partition S by $S = \{S_1, \ldots, S_K\}$, where $S_j \subset \{x_1, \ldots, x_\nu\}$, then the procedure aims to establish an allocation, $\hat{S}$, such that

$$\hat{S} = \text{argmin}_S \left\{ \sum_{i=1}^{M} \sum_{j=1}^{\nu} \mathbb{1}_{\{x_j \in S_i\}} \beta_{x_j} ||x_j - \mu_i||^2 \right\}.$$

The weighted K means algorithm begins with an initial set of K centres $\{\mu_1, \ldots, \mu_K\}$. It then proceeds by alternating between two updating steps until point allocations do not change (signaling a minimum of or a pre-specified number of iterations is reached). A *point allocation step*, where each point, x, is assigned to the set S_j where $j =$

$\arg\min_j ||x-\mu_j||^2$. A *centre point update step* where for the new allocation the centring points are each updated to be the weighted mean of their respective component steps, i.e.

$$\mu_i = \frac{\sum_{j\in S_i} x_j \beta_{x_j}}{\sum_{j\in S_i} \beta_{x_j}}.$$

The Weighted K means procedure, e.g. [15], [44], can be implemented using the R package"FactoClass", by [7] which uses a modified version of the K means algorithm of [13].

6. Theoretical underpinnings of QuanTA in high dimensions

In both QuanTA and the PT algorithms, the acceptance of temperature swap proposals allow the transfer of hot-state mixing information to be passed through to the cold state. The ambitiousness of the spacings between the consecutive inverse temperatures dictate the performance of the algorithm. Similarly to the problem of tuning the RWM algorithm, [31], one seeks the optimal balance between over and under-ambitious proposals. This issue becomes increasingly problematic with an increase in dimensionality, hence careful scaling of the consecutive temperatures spacings is needed to prevent degeneracy of acceptance rates.

The work in [1] sought an optimal scaling result for temperature spacings in a PT algorithm. This section takes a similar approach to derive an equivalent result for QuanTA. It will be shown in Theorem 1 that consecutive spacings inducing swap rates of approximately 0.234 are optimal; thus giving guidance for practitioners to tune towards an optimal setup. Complementary to this, Theorem 2 justifies the use of QuanTA outside the Gaussian setting; showing that under mild conditions the transformation move allows for larger spacings in the temperature schedule than the PT algorithm does.

6.1. Optimal scaling of QuanTA- the setup and assumptions

As the dimensionality, d, of the target distribution tends to infinity, the problem of selecting temperature spacings for QuanTA is investigated. Suppose a swap move between two consecutive temperature levels, β and $\beta' = \beta + \epsilon$ for some $\epsilon > 0$ is proposed. As in [1], the measure the efficiency of the inverse temperature spacing will

be the expected squared jumping distance, $ESJD_\beta$, defined as

$$ESJD_\beta = \mathbb{E}_{\pi_n}\left[(\gamma - \beta)^2\right] \tag{13}$$

where $\gamma = \beta + \epsilon$ if the proposed swap is accepted and $\gamma = \beta$ otherwise. Note the assumption that the Markov chain has reached invariance and so the expectation is taken with respect to the invariant distribution, $\pi_n(\cdot)$.

The $ESJD_\beta$ is a natural quantity to consider, [38], since maximising this would appear to ensure that one is being sufficiently ambitious with spacings but not inducing degenerate acceptance rates. However, it is worth noting that it is only truly justified when there is an associated diffusion limit for the chain, [34].

The aim is to establish the limiting behaviour of the $ESJD_\beta$ as $d \to \infty$ and then optimise this limiting form. To this end, for tractability, the form of the d-dimensional target is restricted to distributions of the form:

$$\pi(x) \propto f_d(x) = \prod_{i=1}^{d} f(x_i). \tag{14}$$

and to achieve a non-degenerate acceptance rate as $d \to \infty$ the spacings are necessarily scaled as $\mathcal{O}(d^{-1/2})$, i.e.

$$\epsilon = \frac{\ell}{d^{1/2}}. \tag{15}$$

where ℓ a positive constant that one tunes to attain an optimal $ESJD_\beta$.

Furthermore, assume that the univariate marginal components, $f(\cdot)$, are C^4 and have a global maxima at μ, i.e.

$$f'(\mu) = 0 \quad \text{and} \quad f(\mu) > f(z)\ \forall z \neq \mu. \tag{16}$$

The point, μ, is the centring point for the transformation aided temperature swap move analysed in the forthcoming Theorem 1. Note that the result of Theorem 1 still holds even if μ is chosen as an arbitrary point in the state space. However, the global mode point, μ, is the canonical transformation centring point as will be seen in the subsequent result in Theorem 2.

Furthermore, the marginal components $f(\cdot)$ are assumed to be of the form

$$f(x) = e^{-H(x)} \qquad \forall x \in \mathbb{R} \tag{17}$$

where the $H(x) := -\log(f(x))$ is regularly varying, [3] i.e. there exists an $\alpha > 0$ such that for $x > 0$

$$\frac{H(tx)}{H(t)} \to x^{\alpha} \quad \text{as} \quad |t| \to \infty. \tag{18}$$

This is a sufficient condition for Theorem 1 and ensures the moments and integrals required for the proof are all well defined. Further assume that the fourth derivatives of $(\log f)(\cdot)$ are bounded, i.e. $\exists M > 0$ such that

$$|(\log f)''''(z)| < M \quad \forall z \in \mathbb{R}. \tag{19}$$

This condition is sufficient for proving Theorem 1 but not necessary. The proof still works if the condition is weakened so that for some $k \geq 4$ then the k^{th} derivative of the logged density is bounded.

Finally, for notational convenience, the following are defined, with the subscript β indicating that the expectation is with respect to $f^{\beta}(\cdot)$:

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

Theorem 1 below deals with a single mode setting which is a very good approximation of a multimodal setting with homogeneous modal structure and where the mode weights remain stable at consecutive temperatures, [48],[40]. Outside this setting other methods are needed to stabilise the modal weights at different temperatures, [41], then the QuanTA transformation-aided swap moves can be successfully utilised alongside such an approach.

Since we assume only one transformation-centring point there is a simplified form of the acceptance probability that no-longer requires the indicator functions. Denote the acceptance probability of the QuanTA-style swap move by $\alpha_{\beta}(x, y)$ so

$$\alpha_{\beta}(x,y) = \min\left(1, \frac{f_d^{\beta'}(g(x,\beta,\beta'))f_d^{\beta}(g(y,\beta',\beta))}{f_d^{\beta'}(y)f_d^{\beta}(x)}\right), \tag{20}$$

then a simple calculation shows that the $ESJD_{\beta}$ from (13), becomes

$$ESJD_{\beta} = \epsilon^2 \mathbb{E}_{\pi_n}\left[\alpha_{\beta}(x,y)\right] \tag{21}$$

which will be maximised with respect to ℓ in the limit as $d \to \infty$.

6.2. Scaling results and interpretation

Under the setting of Section 6.1 and with $\Phi(\cdot)$ denoting the CDF of a standard Gaussian, the following optimal scaling result is derived:

Theorem 1. (Optimal Scaling for the QuanTA Algorithm.) *Consider QuanTA targeting a distribution, $\pi(\cdot)$, satisfying* (14)*. Assume that the marginal components, $f(\cdot)$, satisfy* (16)*, are regularly varying satisfying* (17) *and* (18)*, and that* $\log f(\cdot)$ *satisfies* (19)*. Assuming $\epsilon = \ell/d^{1/2}$ for some $\ell \in \mathbb{R}_+$ then in the limit as $d \to \infty$, the $ESJD_\beta$, given in* (21) *is maximised when ℓ is chosen to maximise*

$$2\ell^2\Phi\left(-\frac{\ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}{\sqrt{2}}\right), \tag{22}$$

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

Proof. The details of the proof of Theorem 1 are deferred to the Appendix, Section A. The strategy comprises 3 key stages which are: establishing a Taylor series expansion of the logged swap move acceptance ratio (i.e. the log of (20)); establishing limiting Gaussianity of this logged acceptance ratio; and finally, achieving a tractable form of the limiting $ESJD_\beta$ which is then optimised with respect to ℓ giving rise to an associated optimal acceptance rate. □

Remark 3.

In the special case that the marginal targets are Gaussian, i.e. $f(x) = \phi(x; \mu, \sigma^2)$ then the transformation swap move should permit arbitrarily ambitious spacings. This is verified by observing that in this case

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

and so with respect to ℓ (22) becomes proportional to ℓ^2 which has no finite maximal value; thus demonstrating consistency with what is know in the Gaussian case.

Remark 4. The optimality criterion given in (22) is very similar to that derived in [1] and [34]. Indeed, both QuanTA and the PT algorithm require the same dimensionality

spacing scaling and both are optimised when a 0.234 acceptance rate is induced. However, there will be a difference in the behaviour of the optimal $\hat{\ell}$ which is where QuanTA can be shown to give accelerated mixing versus the PT approach, see Theorem 2 below.

Remark 5. Theorem 1 gives an explicit formula for derivation of the optimal $\hat{\ell}$ between consecutive temperatures but this is usually intractable in a real problem. However, for a practitioner, the associated 0.234 optimal swap acceptance rate gives useful setup guidelines. In fact, the theorem 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 [30], then [20] took an adaptive MCMC approach, [33], to do this for the PT algorithm but their framework also extends naturally to QuanTA.

6.2.1. *Higher Order Scalings at Cold Temperatures* For any univariate Gaussian distribution at inverse temperature level β, $I(\beta) = 1/(2\beta^2)$. It is shown in [1] that the optimal choice for the scaling parameter takes the form

$$\hat{\ell} \propto I(\beta)^{-1/2} \propto \beta \tag{23}$$

resulting in a geometrically spaced temperature schedule.

Assuming appropriate smoothness for the marginal components , $f(\cdot)$, then for a sufficiently cold temperature the local mode can be well approximated by a Gaussian, [43]. So for sufficiently cold temperatures (i.e., large β) one expects $I(\beta) \approx 1/(2\beta^2)$; thus spacings become (approximately) $\mathcal{O}(\beta)$ (note that a rigorous derivation that $I(\beta) \approx 1/(2\beta^2)$ is contained in the proof of Theorem 2). Defining the "order of the spacing with respect to the inverse temperature, β" as the value of ζ such that the optimal spacing is $\mathcal{O}(\beta^{\zeta})$ then the standard PT algorithm is order 1 for sufficiently cold temperatures.

In the Gaussian setting, QuanTA exhibits "infinitely" high order behaviour since there is no restriction on the size of the temperature spacings with regards the value of β. It is hoped that some of this higher order behaviour is inherited in a more general target distribution setting when the modes can be considered to be sufficiently close to Gaussian.

It will be seen in Theorem 2 below that, under the setting of Theorem 1 but with a single additional condition it is shown that QuanTA does exhibit higher order behaviour than the PT algorithm at cold temperatures.

With $f(\cdot)$ as in Theorem 1 (but now without loss of generality the mode point is at $\mu = 0$) define the normalised density $g_\beta(\cdot)$ as

$$g_\beta(y) \propto f^\beta\left(\mu + \frac{y}{\sqrt{-\beta(\log f)''(\mu)}}\right) = f^\beta\left(\frac{y}{\sqrt{-\beta(\log f)''(0)}}\right). \tag{24}$$

The additional assumption required to prove the higher order behaviour of QuanTA is that there exists $\gamma > 0$ such that as $\beta \to \infty$

$$\left|\mathrm{Var}_{g_\beta}\left(Y^2\right) - 2\right| = \mathcal{O}\left(\frac{1}{\beta^\gamma}\right). \tag{25}$$

This assumption essentially guarantees the convergence to Gaussianity about the mode as $\beta \to \infty$. This assumption appears to be reasonable with studies of both a Gamma and a student-t distributions demonstrating a value of $\gamma = 1$; details can be found in [39].

Theorem 2. (Cold Temperature Scalings.) *For marginal targets, $f(\cdot)$, satisfying the conditions of Theorem 1 and* (25)*, then for β sufficiently large*

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

where

- $k = \min\{2+\gamma, 3\} > 2$ *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

$$\hat{\ell} = \mathcal{O}\left(\beta^{\frac{k}{2}}\right), \tag{26}$$

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

Proof. Since the optimal ℓ derived in Theorem 1 is given by

$$\hat{\ell} \propto \left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{-1/2}$$

the proof of Theorem 2 follows immediately if it can be shown that

$$\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = \mathcal{O}\left(\frac{1}{\beta^k}\right). \tag{27}$$

Indeed, two key Lemmata are derived in Section A.2 in the Appendix: Lemma 4 establishes that $\frac{1}{2}V(\beta) - I(\beta) = \mathcal{O}\left(\frac{1}{\beta^k}\right)$ and Lemma 5 establishes that $\frac{1}{4\beta}R(\beta) = \mathcal{O}\left(\frac{1}{\beta^k}\right)$. Thus the result in (27) holds and the proof is complete.

□

Remark 6. Theorem 2 studies the behaviour of the optimal spacings for large values of β, where β is larger than 1. Even though the QuanTA and PT approaches use temperatures in the range $[0, 1]$, the result still proves insightful in many situations that QuanTA can encounter.

This is due to the arbitrary nature of the "$\beta = 1$" temperature level which is a result of the tempering being relative rather than absolute for the modes. Hence the modes of the target distribution of interest may exhibit a within mode structure that makes it appear that a modes has been cooled to an inverse-temperature much larger than 1.

As an example, take a Gamma(α, γ) distribution, where for some $\alpha > 0$, $\gamma > 0$ and $x > 0$

$$f(x) \propto x^{\alpha-1}e^{-\gamma x}.$$

Then with slight abuse of notation, $f^{\beta^*} \sim$Gamma$(\alpha\beta^* - \beta^* + 1, \gamma\beta^*)$. Suppose that the target distribution of interest was given by $\pi \sim f^{\beta^*}$ and one raises π to a power $\beta < 1$ then $\pi^{\beta} \sim$Gamma$(\alpha\beta\beta^* - \beta\beta^* + 1, \gamma\beta\beta^*)$. Thus if β^* is large and $\beta < 1$ is such that $\beta\beta^*$ remains large then one is in the setting of Theorem 2.

Furthermore, ongoing work is hoping to develop novel methodology for sampling from multimodal distributions that utilises temperature levels with $\beta > 1$ where necessarily $\beta \to \infty$ as $d \to \infty$. The results established here are directly applicable to that work.

Remark 7. The result in Theorem 2 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 the transformation in the transformation-aided swap move approximately preserves the mode's CDF. What Theorem 2 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 colder temperature levels, since it enables accelerated mixing speed through the temperature schedule.

7. Examples of implementation

This section gives illustrative examples for the canonical setting of a Gaussian mixture to illustrate the potential gains of QuanTA over the standard PT approach.

The QuanTA transformation move does not solve all the issues inherent in the PT framework. This will be highlighted with the final example in this section. In fact, [47] and [48] shows that for most "interesting" examples the mixing speed decays exponentially slowly with dimension. Prototype approaches to navigating this problem can be found in [39] and [41]. Also note that the examples all focus on the setting when the target distribution is a Gaussian mixture; which is often a good approximation to a range of multimodal distributions. Outside the setting of Gaussian modes the transformation-aided swap move will not be as effective since the within-mode CDF preservation transformation will only be an approximation to preservation of the CDF.

In each of the examples given, both the new QuanTA and standard (PT) parallel schemes will be run for comparison of performance. In all examples:

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

2. 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. Also all schemes had the same within to swap move ratio $(3 : 1)$.

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

4. Also presented is the optimal temperature schedule for the PT setup generated under the optimal acceptance rate of 0.234 for the PT algorithm suggested by [1]. This demonstrates the extra complexity needed to produce a functioning algorithm for the PT approach.

5. For all runs, the within temperature level proposals were made with Gaussian RWM moves tuned to an optimal 0.234 acceptance rate, [31].

7.1. One-dimensional example

Target distribution given by:

$$\pi(x) \propto \sum_{k=1}^{5} w_k \phi(x; \mu_k, \sigma^2) \tag{28}$$

where $\phi(\cdot; \mu, \sigma^2)$ 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 $w_1 = w_2 = \ldots = w_5$.

The temperature schedule for this example is given by a geometric schedule with an ambitious 0.0002 common ratio for the spacings. Only 3 levels are used and so the temperature schedule is given by $\Delta = \{1, 0.0002, 0.0002^2\}$, see Figure 3.

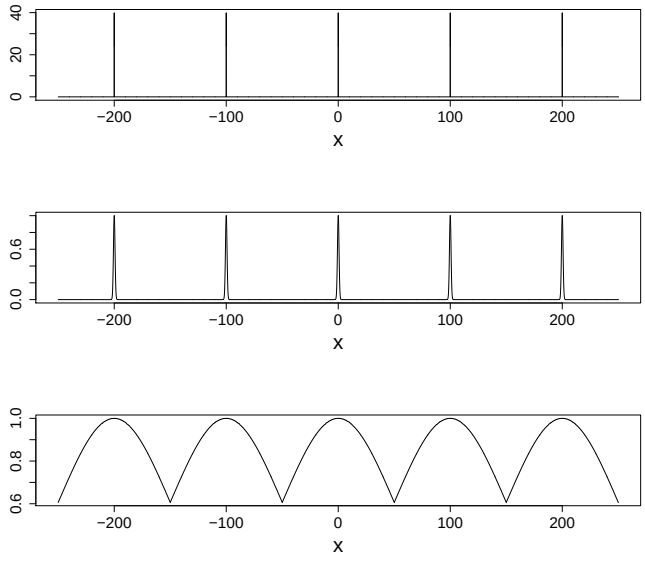

FIGURE 3: The (non-normalised) tempered target distributions for (28) for inverse temperatures $\Delta = \{1, 0.0002, 0.0002^2\}$ respectively.

In all runs all the chains were started from a start location of -200. Figure 1, from the introductory section, shows two representative trace plots of the target state chain

for a run of the PT algorithm and a single scheme from QuanTA respectively. There is a clear improvement in the inter-modal mixing for the QuanTA.

Table 1 gives the associated acceptance rates. Clearly the rate of transfer of mixing information from the hot states to the cold state is significantly higher for QuanTA.

Swap location:	1	2
PT	0.06	0.07
QuanTA	0.99	0.99

TABLE 1: Comparison of the acceptance rates of swap moves for the PT algorithm and QuanTA targeting the one dimensional distribution given in (28) and setup with the ambitious inverse temperature schedule given by $\Delta = \{1, 0.0002, 0.0002^2\}$.

Figure 4 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 10 separate runs of the QuanTA algorithm.

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,

$$\hat{w_k} = \frac{1}{N - B + 1} \sum_{i=B}^{N} \mathbb{1}_{\{c_k < X_i \leq C_k\}}. \tag{29}$$

where c_k and C_k are the chosen upper and lower boundary points for allocation to the k^{th} mode; and B is the length of the burn-in removed.

Figure 4 shows the QuanTA approach has a vastly improved rate of convergence; with the PT runs still exhibiting bias from the chain initialisation locations.

An interesting comparison between the approaches is to observe 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.

With the same hottest state level of $\beta = 0.0002^2$, a geometrical inverse temperature schedule was tuned to give a swap rate of approximately 0.234 was achieved between consecutive levels for the PT algorithm in this example. In fact a 0.04 geometric ratio suggested optimality for the PT scheme. Hence, to reach the stated hottest level needs

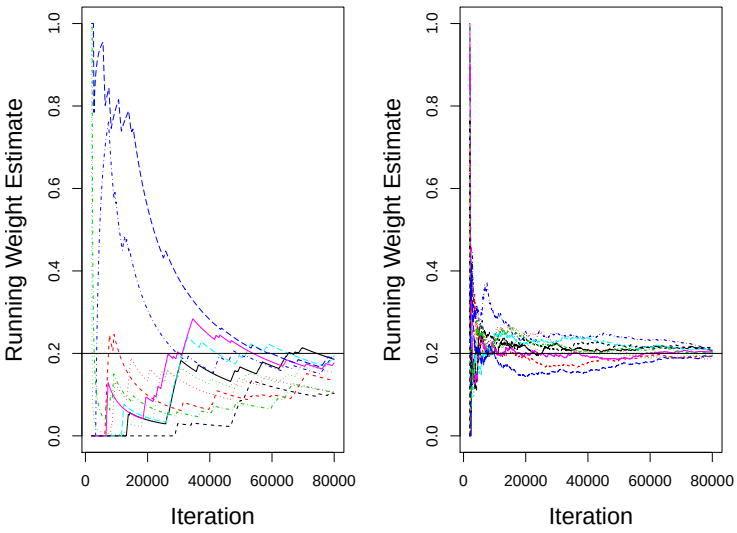

FIGURE 4: For the target given in (28), the running weight approximations for the mode centred on 200 with target weight $w_5 = 0.2$ for 10 separate runs of the PT and QuanTA schemes respectively. Left: the PT runs showing slow and variable estimates for w_5. Right: the new QuanTA scheme showing fast, unbiased convergence to the true value for w_5

7 temperatures, as opposed to the 3 that were evidently unambitious for QuanTA.

7.2. Twenty-dimensional example

The target distribution is a 20-dimensional tri-modal Gaussian:

$$\pi(x) \propto \sum_{k=1}^{3} w_k \left[\prod_{j=1}^{20} \phi(x_j; \mu_k, \sigma^2) \right]. \tag{30}$$

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 $w_1 = w_2 = w_3$.The temperature schedule for this example is derived from a geometric schedule with an ambitious 0.002 common ratio for the spacings. Only 4 levels are used and so the temperature schedule is given by $\{1, 0.002, 0.002^2, 0.002^3\}$.

In all runs all the chains were started from a start location of $(-20, \ldots, -20)$. Figure 5 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

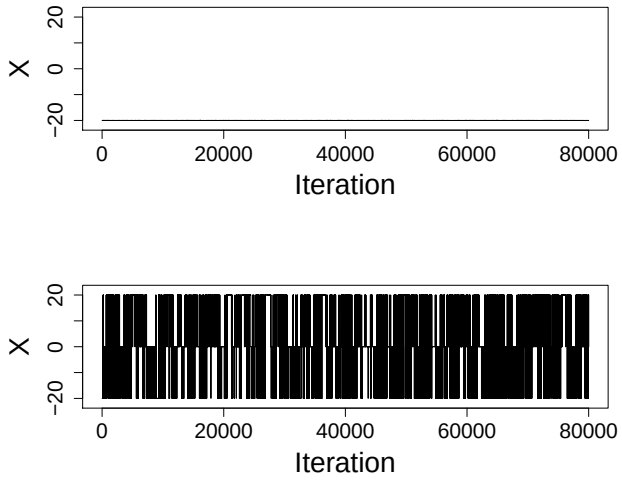

FIGURE 5: Trace plots of the first component of the twenty dimensional cold state chains for representative runs of the PT (top) and new QuanTA (bottom) schemes. Note the fast inter-modal mixing of the new QuanTA scheme, allowing rapid exploration of the target distribution. In contrast the PT scheme never escapes the initialising mode.

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 modes.

The consecutive swap acceptance rates between the four levels are given in Table 2. Clearly there is no transfer of mixing information from the hot states to the cold state for the PT algorithm but there is excellent mixing in the QuanTA.

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

TABLE 2: Comparison of the acceptance rates of swap moves for the PT algorithm and QuanTA targeting the Twenty dimensional distribution given in (30) and setup with the ambitious inverse temperature schedule given by $\{1, 0.002, 0.002^2, 0.002^3\}$.

The temperature schedule choice that induces a 0.234 swap acceptance rate between consecutive temperature levels for this example using the PT algorithm indicates a geometric schedule with a 0.58 common ratio. This is in stark contrast to the 0.002 ratio that is evidently underambitious for QuanTA. Indeed, to reach the allocated hot state of $\beta = 0.002^3$ then the PT algorithm would need 36 temperature levels in contrast to the 4 that sufficed for QuanTA.

7.3. Five-dimensional non-canonical example

Leaving the canonical symmetric mode 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:

$$\pi(x) \propto \sum_{k=1}^{3} w_k \left[\prod_{j=1}^{5} \phi(x_j; \mu_k, \sigma_k^2) \right]. \tag{31}$$

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 does not 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, [48]. Indeed, [41] and [40] present a novel modification to the auxiliary tempered targets that remain compatible with the QuanTA scheme.

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:

$$\Delta = \{1, 0.08, 0.08^2, 0.08^3, 0.4^9, 0.4^{10}, \ldots, 0.4^{15}, 0.4^{16}\}. \tag{32}$$

For the QuanTA scheme, the transformation moves were used for swap moves between

the coldest 7 levels and standard swap moves were used otherwise.

Figure 6 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. 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.

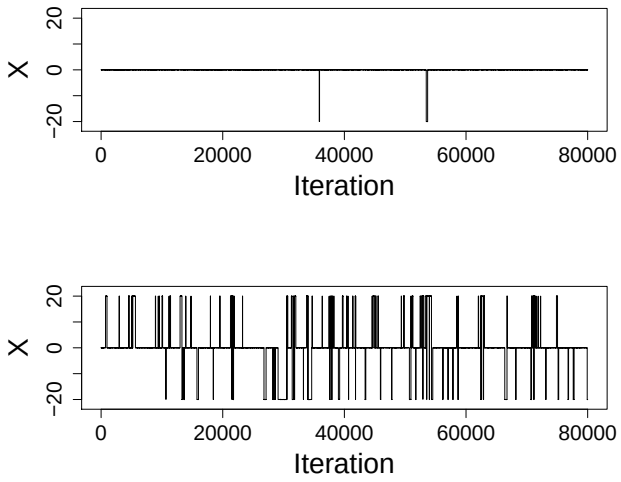

FIGURE 6: Trace plots of the first component of the five dimensional cold state chains for representative runs of the PT and QuanTA schemes respectively. Note the difference in inter-modal mixing between the QuanTA scheme and the PT scheme which struggles to escape the initialisation mode.

The consecutive swap acceptance rates between the 12 levels are given in Table 2. Clearly there is very poor mixing through the 4 coldest states for the PT algorithm. In contrast the QuanTA scheme has solid swap acceptance rates through the coldest levels but, unlike the previous examples, they are not all close to 1.

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

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 3: Comparison of the acceptance rates of swap moves for the PT and new QuanTA algorithm targeting the five dimensional distribution given in (31) and setup with the ambitious inverse temperature schedule given in (32). Note that for QuanTA, the reparametrised swap move was only used for swaps in the coldest 7 levels.

PT and new QuanTA schemes when using power-based tempering on a target outside of the canonical symmetric mode setting.

7.4. The computational cost of QuanTA

It is important to analyse the computational cost of QuanTA. 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.

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

TABLE 4: Complexity comparisons between QuanTA and PT for the one-dimensional example.

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

TABLE 5: Complexity comparisons between QuanTA and PT for the twenty-dimensional example.

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 fundamental value for comparison. 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 fully fair since it is sharing out the clustering expense between schemes.
- 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 effective

sample size 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.

8. Conclusion and further work

The prototype QuanTA approach utilises a non-centred transformation approach to accelerate the transfer of mixing information from the rapidly mixing "hot" state to aid the inter-modal mixing in the target "cold" state. Examples show that this novel algorithm has the potential to dramatically improve the inferential gains; particularly in settings where the modes are similar to a Gaussian in structure.

The accompanying theoretical results that are given in Section 6 show that in a generic non-Gaussian setting the QuanTA approach can still exhibit accelerated mixing through the temperature schedule. Although the inverse temperature spacings are generally still $\mathcal{O}(d^{-1/2})$ there is a higher order behaviour exhibitted in the mixing for large (i.e. cold) values of the inverse temperature β. This suggests that the QuanTA approach will be powerful for accelerating the mixing through the colder levels of the temperature schedule for a typical smooth target.

It is clear that there are interesting questions to be addressed and further work needed before QuanTA can be considered practical in a real data problem. In terms of optimising the computational expense, it has been shown that parallelisation of the PT algorithm can give significant practical gains, [45]; by design QuanTA is also highly parallelisible. The procedure described in Section 4.1 has huge scope to be parallelised. This requires care since communication costs and synchronisation between parallel collections of the PT schemes could negate the effectiveness of the parallelisation.

A criticism of the current clustering method used is that it requires prior specification of the number of modes, K which is likely to be unknown and would need online-estimation as part of the clustering process.

An interesting question is whether using extra cold levels (with $\beta > 1$) along with the weighted clustering would help to aid the stability of the clustering procedure once invariance is reached for the population. Indeed, the mixing at these auxiliary extra

cold levels should be very fast due to the QuanTA exhibiting higher order behaviour in these modes. The other interesting question is regarding the robustness of the method outside of Gaussian modes e.g., in heavier tailed modes, when the Gaussian approximation to the mode can be poor. Consider the setting of a univariate Laplace distribution and observe that the QuanTA style transformation never agrees with the true CDF preserving transformation. It would be interesting to search, both empirically and theoretically, for settings where QuanTA fails to outperform the standard PT swap move. Some initial ideas and details of this further work can be found in [39].

Appendix A. Proofs

This section gives the proof details of the results in Section 6.2. Firstly, some key notation is introduced that will be useful throughout this section.

Definition 1. Denote:

- $B = \log\left(\frac{f_d^{\beta'}(t(x,\beta,\beta'))f_d^{\beta}(t(y,\beta',\beta))}{f_d^{\beta'}(y)f_d^{\beta}(x)}\right)$;
- $h(x) := \log\left(f(x)\right)$;
- $k(x) := (x-\mu)h'(x)$;
- and $r(x) := (x-\mu)^2 h''(x)$.

Then define

$$M(\beta) = \mathbb{E}_\beta(h(z)) \tag{33}$$

$$S(\beta) = \mathbb{E}_\beta(k(z)) \tag{34}$$

$$R(\beta) = \mathbb{E}_\beta(r(z)-k(z)), \tag{35}$$

where all expectations are with respect to the distribution $\frac{f^\beta(x)}{Z_\beta}$ where $Z_\beta = \int f^\beta(z)dz$.

Proposition 2. *Under the notation and assumptions of Theorem 1 and definition 1 then it can be shown that*

$$I(\beta) := M'(\beta) = \mathrm{Var}_\beta(h(x)). \tag{36}$$

and

$$S(\beta) = -\frac{1}{\beta}, \tag{37}$$

which trivially gives that

$$V(\beta) := S'(\beta) = \frac{1}{\beta^2}. \tag{38}$$

Proof. The proof of (36) is routine and can be found in [1]. The derivation of (37) is less obvious using integration by parts:

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

□

A.1. Proof of Theorem 1

This section derives 3 key results that are specific to deriving the result in Theorem 1. Lemma 1 will establish a Taylor expanded form of the log acceptance ratio of a temperature swap move that will prove to be asymptotically useful. Lemma 2 will then establish the limiting Gaussianity of this logged acceptance ratio and finally, Lemma 3 completes the proof of Theorem 1 by establishing the optimal spacings and associated optimal acceptance rates required.

Lemma 1. (QuanTA Log-Acceptance Ratio.) *Under the notation and assumptions of Theorem 1 and definition 1,*

$$\begin{aligned} B &= \epsilon\left[\sum_{i=1}^{d} h(x_i) - h(y_i) + \frac{1}{2}\left(k(y_i) - k(x_i)\right)\right] \\ &\quad + \frac{\epsilon^2}{8\beta}\left[\sum_{i=1}^{d} r(x_i) - k(x_i) + r(y_i) - k(y_i)\right] + (T_x + T_y). \end{aligned} \tag{39}$$

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

Proof. By taking logarithms it is immediate that

$$\begin{aligned} B &= \sum_{i=1}^{d}[\beta' h(t(x_i, \beta, \beta^{'})) - \beta h(x_i)] + \sum_{i=1}^{d}[\beta h(t(y_i, \beta^{'}, \beta)) - \beta' h(y_i)] \\ &=: H_\beta^{\beta'}(\mathbf{x}) + H_{\beta'}^{\beta}(\mathbf{y}). \end{aligned} \tag{40}$$

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 $h(t(x,\beta,\beta^{'}))$ explicitly dependent on ϵ

$$\alpha_x(\epsilon) := h(t(x,\beta,\beta^{'})) = \log\left[f\left(\left(\frac{\beta}{\beta+\epsilon}\right)^{1/2}(x-\mu)+\mu\right)\right].$$

- 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 $0<\xi_x<\epsilon$:

$$h(g(x,\beta,\beta^{'})) = \alpha_x(\epsilon) = \alpha_x(0)+\epsilon\alpha_x^{'}(0)+\frac{\epsilon^2}{2}\alpha_x^{''}(0)+\frac{\epsilon^3}{6}\alpha_x^{'''}(\xi_x), \tag{41}$$

where

$$\begin{aligned}
\alpha_x^{'}(\epsilon) &= -\frac{(x-\mu)}{2}\frac{\beta^{1/2}}{(\beta+\epsilon)^{3/2}}(\log f)^{'}(d_x(\epsilon)), &(42)\\
\alpha_x^{''}(\epsilon) &= \frac{(x-\mu)^2}{4}\frac{\beta}{(\beta+\epsilon)^3}(\log f)^{''}(d_x(\epsilon)) &(43)\\
&\quad +\frac{3(x-\mu)}{4}\frac{\beta^{1/2}}{(\beta+\epsilon)^{5/2}}(\log f)^{'}(d_x(\epsilon)),\\
\alpha_x^{'''}(\epsilon) &= -\frac{(x-\mu)^3}{8}\frac{\beta^{3/2}}{(\beta+\epsilon)^{9/2}}(\log f)^{'''}(d_x(\epsilon)) &(44)\\
&\quad -\frac{9(x-\mu)^2}{8}\frac{\beta}{(\beta+\epsilon)^4}(\log f)^{''}(d_x(\epsilon))\\
&\quad -\frac{15(x-\mu)}{8}\frac{\beta^{1/2}}{(\beta+\epsilon)^{7/2}}(\log f)^{'}(d_x(\epsilon)).
\end{aligned}$$

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 $0 < |\xi_k^\epsilon - x| < |d_x(\epsilon) - x|$

$$\begin{aligned}
(\log f)'(d_x(\epsilon)) &= (\log f)'(x) + \eta_\epsilon (x-\mu)(\log f)''(x) \\
&\quad + \frac{{\eta_\epsilon}^2 (x-\mu)^2}{2} (\log f)'''(x) \\
&\quad + \frac{{\eta_\epsilon}^3 (x-\mu)^3}{6} (\log f)''''(\xi_1^\epsilon), \qquad (45) \\
(\log f)''(d_x(\epsilon)) &= (\log f)''(x) + \eta_\epsilon (x-\mu)(\log f)'''(x) \\
&\quad + \frac{{\eta_\epsilon}^2 (x-\mu)^2}{2} (\log f)''''(\xi_2^\epsilon), \qquad (46) \\
(\log f)'''(d_x(\epsilon)) &= (\log f)'''(x) + \eta_\epsilon (x-\mu)(\log f)''''(\xi_3^\epsilon). \qquad (47)
\end{aligned}$$

Recall the assumptions (18) and (19). Substituting (45), (46) and (47) into (44); evaluating the expectation with respect to $X \sim f^\beta$ and for convenience denoting $|x-\mu|$

by S then $\exists\, C \in \mathbb{R}_+$

$$\begin{aligned}
|\mathbb{E}_\beta[\alpha'''_x(\xi_x)]| &\leq \mathbb{E}_\beta\left[|\alpha'''_x(\xi_x)|\right] \\
&\leq \mathbb{E}_\beta\Bigg[\frac{S^3}{8}\beta^{-3}|(\log f)'''(d(\xi_x))| \\
&\quad + \frac{9S^2}{8}\beta^{-3}|(\log f)''(d(\xi_x))| + \frac{15S}{8}\beta^{-3}|(\log f)'(d(\xi_x))|\Bigg] \\
&\leq \mathbb{E}_\beta\Bigg[\frac{S^3}{8}\beta^{-3}\left(|(\log f)'''(x)| + S|(\log f)''''(\xi_3^{\xi_x})|\right) \\
&\quad + \frac{9S^2}{8}\beta^{-3}\Bigg(|(\log f)''(x)| + S|(\log f)'''(x)| \\
&\quad + \frac{|x|^2}{2}|(\log f)''''(\xi_2^{\xi_x})|\Bigg) + \frac{15S}{8}\beta^{-3}\Bigg(|(\log f)'(x)| \\
&\quad + S|(\log f)''(x)| + \frac{S^2}{2}|(\log f)'''(x)| \\
&\quad + \frac{S^3}{6}|(\log f)''''(\xi_1^{\xi_x})|\Bigg)\Bigg] \leq C \qquad (48)
\end{aligned}$$

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 (41), with substitution of terms from (42), (43) and (44), $H_\beta^{\beta'}(\mathbf{x})$ can be expressed as

$$\begin{aligned}
H_\beta^{\beta'}(\mathbf{x}) &= \sum_{i=1}^{d}(\beta+\epsilon)\left[\alpha_{x_i}(\epsilon) - \beta\alpha_{x_i}(0)\right] \\
&= \epsilon\sum_{i=1}^{d}\left[\alpha_{x_i}(0) + \beta\alpha'_{x_i}(0)\right] + \epsilon^2\sum_{i=1}^{d}\left[\frac{\beta}{2}\alpha''_{x_i}(0) + \alpha'_{x_i}(0)\right] \\
&\quad + \epsilon^3\sum_{i=1}^{d}\left[\frac{1}{2}\alpha''_{x_i}(0) + \frac{\beta}{6}\alpha'''_{x_i}(\xi_{x_i})\right] + \epsilon^4\sum_{i=1}^{d}\frac{1}{6}\alpha'''_{x_i}(\xi_{x_i}). \qquad (49)
\end{aligned}$$

By (48) and using the iid nature of the $x_i's$ and using Markov's inequality then $\forall \delta > 0$

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

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,

$$\begin{aligned}H_\beta^{\beta'}(\mathbf{x}) \quad = \quad & \epsilon\left[\sum_{i=1}^{d}h(x_i)-\frac{1}{2}(x_i-\mu)h'(x_i)\right] \\ & +\frac{\epsilon^2}{8\beta}\left[\sum_{i=1}^{d}(x_i-\mu)^2h''(x_i)-(x_i-\mu)h'(x_i)\right]+T_x \qquad (50)\end{aligned}$$

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 $h(t(y,\beta^{'},\beta))$ as

$$\alpha_y(\epsilon):=h(t(y,\beta^{'},\beta))=\log\left[f\left(\left(\frac{\beta+\epsilon}{\beta}\right)^{1/2}(y-\mu)+\mu\right)\right],$$

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

$$h(t(y,\beta^{'},\beta))=\alpha_y(\epsilon)=\alpha_y(0)+\epsilon\alpha'_y(0)+\frac{\epsilon^2}{2}\alpha''_y(0)+\frac{\epsilon^3}{6}\alpha_y(\xi_y). \qquad (51)$$

By identical methodology to the above calculation in (48) for $\alpha_x(\cdot)$, it can be shown that $\exists\ C_y\in\mathbb{R}_+$ such that

$$\left|\mathbb{E}_\beta[\alpha'''_y(\xi_y)]\right| \quad \leq \quad C_y. \qquad (52)$$

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

$$\begin{aligned} H^{\beta}_{\beta'}(\mathbf{y}) &= -\epsilon\left[\sum_{i=1}^{d} h(y_i) - \frac{1}{2}(y_i-\mu)h'(y_i)\right] \\ &\quad + \frac{\epsilon^2}{8\beta}\left[\sum_{i=1}^{d}(y_i-\mu)^2 h''(y_i) - (y_i-\mu)h'(y_i)\right] + T_y. \end{aligned} \tag{53}$$

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

Using the notation from Definition 1 the desired form of B in Lemma 1 is reached. □

Lemma 2. (Asymptotic Gaussianity of the Log-Acceptance Ratio for QuanTA.) *Under the notation and assumptions of Theorem 1 and Definition 1, B is asymptotically Gaussian of the form $B \dot{\sim} N(\frac{-\sigma^2}{2}, \sigma^2)$ where*

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

Proof. Recall the form of B from Lemma 1, then 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. [6], for a sum of iid variables, then asymptotic Gaussianity is immediate where

$$W(d) \Rightarrow N\left(\mu_W, \sigma^2_W\right) \quad \text{as} \quad d \to \infty \tag{54}$$

where

$$\mu_W = \lim_{d\to\infty} \mathbb{E}[W(d)] \quad \text{and} \quad \sigma^2_W = \lim_{d\to\infty} \mathrm{Var}[W(d)].$$

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

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

where the final line utilises the results in Proposition 2. 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

$$B \;\dot{\sim}\; N\left(\ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right], 2\ell^2\mathrm{Var}_\beta\left(h(x) - \frac{1}{2}k(x)\right)\right). \tag{55}$$

However, this does not obviously have the form required with $B\dot{\sim}N(\frac{-\sigma^2}{2}, \sigma^2)$ for some σ^2. This form is verified with the following Proposition 3 and this then completes the proof of Lemma 2.

Proposition 3. *Under the notation and assumptions of Theorem 1 and Definition 1 then*

$$\ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] = -\ell^2\mathrm{Var}_\beta\left(h(x) - \frac{1}{2}k(x)\right). \tag{56}$$

Proof. From (55) then denote

$$\mu = \ell^2\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right] \tag{57}$$

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

$$-\frac{\sigma^2}{2} = \ell^2\left[-I(\beta) - \frac{1}{4}\mathrm{Var}_\beta(k(x)) + V(\beta)\right]. \tag{58}$$

Consequently, equating the terms on the RHS of (57) and (58) shows that if the following can be shown to hold then the required identity in (56) is validated:

$$\frac{1}{4\beta}R(\beta) = -\frac{1}{4}\mathrm{var}_\beta(k(x)) + \frac{1}{2}V(\beta). \tag{59}$$

The LHS and RHS of (59) 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 (37) $\mathbb{E}_\beta(k(x)) = -1/\beta$:

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

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

$$\begin{aligned} \mathbb{E}_\beta(k(x)^2) &= \int (x-\mu)^2[(\log f)'(x)]^2 \frac{f^\beta(x)}{Z_\beta}dx \\ &= \left[\frac{(x-\mu)^2}{\beta}(\log f)'(x)\frac{f^\beta(x)}{Z_\beta}\right]_{-\infty}^{\infty} \\ &\quad -\frac{1}{\beta}\int \left[(x-\mu)^2(\log f)''(x) + 2(x-\mu)(\log f)'(x)\right]\frac{f^\beta(x)}{Z_\beta}dx \\ &= 0 - \frac{1}{\beta}\mathbb{E}_\beta(r(x)) - \frac{2}{\beta}\mathbb{E}_\beta(k(x)) = -\frac{1}{\beta}\mathbb{E}_\beta(r(x)) + \frac{2}{\beta^2}. \end{aligned} \tag{60}$$

Collating the above in (59) and (60) then

$$-\frac{1}{4}\text{var}_\beta(k(x)) + \frac{1}{2}V(\beta) = \frac{1}{4\beta}\mathbb{E}_\beta(r(x)) + \frac{1}{4\beta^2} = \frac{1}{4\beta}R(\beta), \tag{61}$$

where the final equality simply comes from the definition of $R(\beta)$ from(35). □

□

Lemma 3. (Optimisation of the $ESJD_\beta$.) *Under the notation and assumptions of Theorem 1 and Definition 1 then the $ESJD_\beta$, is maximised when ℓ is chosen to maximise*

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

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

Proof. Letting $\phi_{(m,\sigma^2)}$ denote the density function of a Gaussian with mean m and variance σ^2 and suppose that $G \sim N(-\frac{\sigma^2}{2}, \sigma^2)$ then a routine calculation (which can be found in e.g. [31]) shows that

$$\mathbb{E}(1 \wedge e^G) = 2\Phi\left(-\frac{\sigma}{2}\right). \tag{62}$$

Using the result in (62) and Lemma 2, then in the limit as $d \to \infty$

$$\lim_{d\to\infty} (d\ ESJD_\beta) = 2\ell^2\Phi\left(-\frac{\ell\left[\frac{1}{2}V(\beta) - I(\beta) + \frac{1}{4\beta}R(\beta)\right]^{1/2}}{\sqrt{2}}\right). \tag{63}$$

Substituting

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

and then maximising with respect to u attains an optimising value u^* that doesn't depend on

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

Recalling the form of the $ESJD_\beta$ from (21) , then it is clear that the associated acceptance rate, denoted (ACC_β), induced by choosing the any value of u is

$$ACC_\beta = \mathbb{E}_{\pi_n}\left[\alpha_\beta(x, y)\right]$$

which, as established above, in the limit as $d \to \infty$ is asymptotically given by

$$ACC_\beta = 2\Phi\left(-\frac{u}{\sqrt{2}}\right)$$

Now it can be shown numerically that for the optimising value u^* induces

$$\mathrm{ACC}_\beta = 0.234 \quad (3.s.f).$$

□

A.2. Proof of Theorem 2

Note that in Theorem 2, the conditions on $f(\cdot)$ are inherited from the conditions on $f(\cdot)$ from Theorem 1. This includes the bounded fourth derivatives of $\log(f)$ and the existence of eighth moments i.e. $\mathbb{E}_\beta\left[X^8\right]$, which is due to the assumption of regularly varying tails. These will be assumed for the following lemmata.

Lemma 4. *Under the notation and assumptions of Theorems 1 and 2 and Definition 1 then*

$$\frac{1}{2}V(\beta) - I(\beta) = \mathcal{O}\left(\beta^{-k}\right)$$

where in general $k = \min\{2 + \gamma, 5/2\}$ *but if* $h'''(0) = 0$ *then* $k = \min\{2 + \gamma, 3\}$.

Proof. 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

$$\begin{aligned} I(\beta) &= \mathrm{Var}_\beta\left[h(X)\right] \\ &= \int \left(h(x) - \mathbb{E}_\beta[h(X)]\right)^2 \frac{f^\beta(x)}{Z(\beta)} dx \\ &= \int \left(h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) - \mathbb{E}_{g_\beta}\left[h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right)\right]\right)^2 g_\beta(y) dy \end{aligned} \quad (64)$$

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

$$\begin{aligned} h\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) &= h(0) - \frac{y^2}{2\beta} \\ &\quad + \frac{y^3 h'''(0)}{6\left(\beta(-h''(0))\right)^{3/2}} + \frac{y^4 h''''(\xi_1(y))}{24\left(\beta(-h''(0))\right)^2} \end{aligned} \quad (65)$$

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 the assumption of bounded fourth derivatives

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

where $\mathbb{E}_{g_\beta}\left[Y^4\right] < \infty$ due to the assumption on the existence of moments up to the eighth moment. Thus,

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

and substituting this into (64), along with the Taylor expansion of h to the fourth

order given in (65), gives

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

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

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

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

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

and so under the key assumption given in (25), then

$$I(\beta) = \frac{1}{2\beta^2} + \mathcal{O}\left(\frac{1}{\beta^k}\right) \tag{66}$$

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

□

Lemma 5. *Under the notation and assumptions of Theorems 1 and 2 and Definition 1 then*

$$\frac{1}{4\beta}R(\beta) = \mathcal{O}\left(\beta^{-k}\right)$$

where in general $k = 5/2$ *but if* $h'''(0) = 0$ *then* $k = 3$.

Proof. Recall that

$$\begin{aligned}\frac{1}{4\beta}R(\beta) &= \frac{1}{4\beta}\mathbb{E}_\beta\left[X^2h''(X) - Xh'(X)\right] \\ &= \frac{1}{4\beta}\mathbb{E}_{g_\beta}\left[\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)^2 h''\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)\right. \\ &\qquad \left. -\frac{Y}{\sqrt{\beta(-h''(0))}}h'\left(\frac{Y}{\sqrt{\beta(-h''(0))}}\right)\right]. \end{aligned} \tag{67}$$

Using Taylor expansion about the mode at 0 then

$$\begin{aligned}h'\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) &= h'(0) + \frac{y}{\sqrt{\beta(-h''(0))}}h''(0) + \frac{y^2}{2\beta(-h''(0))}h'''(0) \\ &\qquad + \frac{y^3}{6\beta^{3/2}(-h''(0))^{3/2}}h''''(\xi_2(y)), \end{aligned} \tag{68}$$

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,

$$\begin{aligned}h''\left(\frac{y}{\sqrt{\beta(-h''(0))}}\right) &= h''(0) + \frac{y}{\sqrt{\beta(-h''(0))}}h'''(0) \\ &\qquad + \frac{y^2}{2\beta^{3/2}(-h''(0))^{3/2}}h''''(\xi_3(y)) \end{aligned} \tag{69}$$

where $\xi_3(\cdot)$ is the truncation term for the Taylor expansion such that $0 < |\xi_3(y)| < \left|\frac{y}{\sqrt{\beta(-h''(0))}}\right|$ for all y. Hence,

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

Substituting this in to the $\frac{1}{4\beta}R(\beta)$ term in (67)

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

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 the assumptions of boundedness of the fourth derivatives of $\log f(X)$ and the existence of moments. Hence, in general

$$\frac{1}{4\beta}R(\beta) = \mathcal{O}\left(\frac{1}{\beta^{5/2}}\right)$$

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

$$\frac{1}{4\beta}R(\beta) = \mathcal{O}\left(\frac{1}{\beta^{3}}\right).$$

Consequently,

$$\frac{1}{4\beta}R(\beta) = \mathcal{O}\left(\frac{1}{\beta^{k}}\right) \tag{70}$$

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

Acknowledgements

Both authors would like to acknowledge the funding from the grants from EPSRC (EP/L505110/1 and EP/KO14463/1) that made this work possible. We would also like to thank both Professors Jonathan Tawn (Lancaster) and Jeff Rosenthal (Toronto) for the helpful suggestions and support during the work.

References

[1] 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.

[2] Barndorff-Nielsen, O. E. and Nielsen, O. E. B. (1989). Asymptotic Techniques; for use in Statistics. *Technical report.*

[3] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989). *Regular variation* vol. 27. Cambridge university press.

[4] 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.

[5] Carter, J. and White, D. (2013). History matching on the imperial college fault model using parallel tempering. *Computational Geosciences* **17,** 43–65.

[6] Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge university press.

[7] 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.

[8] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning* vol. 1. Springer Series in Statistics New York.

[9] Ge, R., Lee, H. and Risteski, A. (2017). Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo.

[10] Geyer, C. J. (1991). Markov chain Monte Carlo Maximum Likelihood. *Computing Science and Statistics* **23,** 156–163.

[11] Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90,** 909–920.

[12] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82,** 711–732.

[13] Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means Custering Algorithm. *Applied statistics* 100–108.

[14] Hastie, D. (2005). Towards Automatic Reversible Jump Markov chain Monte Carlo. *PhD thesis*. University of Bristol.

[15] Kerdprasop, K., Kerdprasop, N. and Sattayatham, P. (2005). Weighted K-means for Density-biased Clustering. In *Data Warehousing and Knowledge Discovery*. Springer pp. 488–497.

[16] Kim, S., Tadesse, M. G. and Vannucci, M. (2006). Variable Selection in Clustering via Dirichlet process Mixture Models. *Biometrika* **93,** 877–893.

[17] 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.

[18] Mangoubi, O., Pillai, N. S. and Smith, A. (2018). Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities?

[19] Marinari, E. and Parisi, G. (1992). Simulated Tempering: a New Monte Carlo Scheme. *EPL (Europhysics Letters)* **19,** 451.

[20] Miasojedow, B., Moulines, E. and Vihola, M. (2013). An Adaptive Parallel Tempering Algorithm. *Journal of Computational and Graphical Statistics* **22,** 649–664.

[21] Mohamed, L., Calderhead, B., Filippone, M., Christie, M. and Girolami, M. (2012). Population MCMC methods for history matching and uncertainty quantification. *Computational Geosciences* **16,** 423–436.

[22] Neal, R. M. (1996). Sampling from Multimodal Distributions using Tempered Transitions. *Statistics and Computing* **6,** 353–366.

[23] Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

[24] Nemeth, C., Lindsten, F., Filippone, M. and Hensman, J. (2017). Pseudo-extended Markov Chain Monte Carlo. *ArXiv e-prints*.

[25] Olver, F. (1968). Error bounds for the Laplace Approximation for Definite Integrals. *Journal of Approximation Theory* **1,** 293–313.

[26] 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*. Oxford University Press, USA. pp. 307–326.

[27] Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science* 59–73.

[28] Paulin, D., Jasra, A., Thiery, A. et al. (2019). Error bounds for sequential Monte Carlo samplers for multimodal distributions. *Bernoulli* **25,** 310–340.

[29] 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.

[30] Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics* 400–407.

[31] 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.

[32] Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and Ergodicity of Adaptive Markov chain Monte Carlo Algorithms. *Journal of Applied Probability* 458–475.

[33] Roberts, G. O. and Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics* **18,** 349–367.

[34] 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.

[35] Roberts, G. O., Rosenthal, J. S. et al. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science* **16,** 351–367.

[36] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)* **71,** 319–392.

[37] Schweizer, N. (2012). Non-asymptotic error bounds for Sequential MCMC methods. *PhD thesis*. Universitäts-und Landesbibliothek Bonn.

[38] Sherlock, C. (2006). Methodology for Inference on the Markov Modulated Poisson Process and Theory for Optimal Scaling of the Random Walk Metropolis. *PhD thesis*. Lancaster University.

[39] Tawn, N. (2017). Towards Optimality of the Parallel Tempering Algorithm. *PhD thesis*. University of Warwick.

[40] Tawn, N. G. and Roberts, G. O. (2018). Optimal temperature spacing for regionally weight-preserving tempering.

[41] Tawn, N. G., Roberts, G. O. and Rosenthal, J. S. (2018). Weight preserving simulated tempering. *Accepted, to appear in Statistics and Computing*.

[42] Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics* 1701–1728.

[43] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81,** 82–86.

[44] Tseng, G. C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* **23,** 2247–2255.

[45] VanDerwerken, D. N. and Schmidler, S. C. (2013). Parallel Markov Chain Monte Carlo. *arXiv preprint arXiv:1312.7479*.

[46] Wang, J.-S. and Swendsen, R. H. (1990). Cluster Monte Carlo algorithms. *Physica A: Statistical Mechanics and its Applications* **167,** 565–579.

[47] Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for Rapid Mixing of Parallel and Simulated Tempering on Multimodal Distributions. *The Annals of Applied Probability* 617–640.

[48] Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering. *Electronic Journal of Probability* **14,** 780–804.

[49] Xie, Y., Zhou, J. and Jiang, S. (2010). Parallel tempering Monte Carlo simulations of lysozyme orientation on charged surfaces. *The Journal of Chemical Physics* **132,** 02B602.