Dynamic temperature selection for parallel-tempering in Markov chain Monte Carlo simulations

W. D. Vousden*, W. M. Farr* and I. Mandel*

University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

26 January 2015

ABSTRACT

Modern problems in astronomical Bayesian inference require efficient methods for sampling from complex, high-dimensional, often multi-modal probability distributions. Most popular methods, such as Markov chain Monte Carlo sampling, perform poorly on strongly multi-modal probability distributions, rarely jumping between modes or settling on just one mode without finding others. Parallel tempering addresses this problem by sampling simultaneously with separate Markov chains from tempered versions of the target distribution with reduced contrast levels. Gaps between modes can be traversed at higher temperatures, while individual modes can be efficiently explored at lower temperatures. In this paper, we investigate how one might choose the ladder of temperatures to achieve lower autocorrelation time for the sampler (and therefore more efficient sampling). In particular, we present a simple, easily-implemented algorithm for dynamically adapting the temperature configuration of a sampler while sampling in order to maximise its efficiency. This algorithm dynamically adjusts the temperature spacing to achieve a uniform rate of exchanges between neighbouring temperatures. We compare the algorithm to conventional geometric temperature configurations on a number of test distributions, and report efficiency gains by a factor of 1.2–2.5 over a well-chosen geometric temperature configuration and by a factor of 1.5–5 over a poorly chosen configuration. On all of these test distributions a sampler using the dynamical adaptations to achieve uniform acceptance ratios between neighbouring chains outperforms one that does not.

1 INTRODUCTION

Many problems in astronomical data analysis and Bayesian statistical inference demand the characterisation of high-dimensional probability distributions with complicated structures. Lacking analytic forms, these distributions must be explored numerically, usually via Monte Carlo methods.

Parallel tempered Markov chain Monte Carlo (MCMC), a development on standard MCMC, uses several Markov chains in parallel to explore a target distribution at different “temperatures” (Earl & Deem 2005; Swendsen & Wang 1986; Geyer 1991). As the temperature increases, the posterior distribution asymptotes to the prior, allowing a chain to efficiently explore the whole prior volume without becoming stuck in regions of the parameter space with high probability density. At lower temperatures, a chain can more efficiently sample from such a high-probability region. Meanwhile, exchange of positions between chains allows colder chains to migrate between widely separated modes in the parameter space (Geyer 1991). Parallel tempered MCMC samplers are thus particularly well-suited to sampling posterior distributions with well-separated modes, where a regular MCMC sampler would take many iterations to find its way between modes.

An open problem in the application of parallel tempering is selecting a specification, or ladder, of temperatures that minimises the autocorrelation time (ACT) of the chain sampling the posterior distribution of interest. The efficiency of a given ladder hinges critically on the rate at which it can transfer the positions in parameter space of samples between high and low temperatures.

In this paper we present a simple algorithm that adapts the temperature ladder of a parallel tempered MCMC sampler such that the rate of exchange between chains is uniform over the entire ladder. The algorithm is easy to implement in existing MCMC code, and we pro-
provide an example implementation for the *emcee* sampler of Foreman-Mackey et al. (2013).

In §3 we describe the parallel tempering formalism and lay out the requirements for a good temperature ladder. We discuss previous work on temperature selection and suggest a definition of ladder optimality that, for simple cases, proposes a geometric spacing of temperatures. For illustration, we apply these ideas in §2.2 to the simple example of an unbounded Gaussian posterior distribution.

In §4 we describe the algorithm mentioned above and then apply it in §1 to a variety of test distributions. We show that, while our temperature selection strategy is not necessarily optimal in the ACT of the sampler, it nonetheless improves the ACT compared to the simple geometric spacing that is conventional in the literature (Earl & Deem 2005; Sugita & Okamoto 1999; Kolpe 2002, 2004) by factors of >1.2 for our test cases.

We conclude in §3 and §5 with a discussion of our results and suggestions for further research.

2 PARALLEL TEMPERING

Parallel tempering (Earl & Deem 2005; Swendsen & Wang 1986; Geyer 1991) is a development on the standard MCMC formalism that uses several Markov chains in parallel to sample from tempered versions of the posterior distribution $\pi$,

$$\pi_T(\vec{\theta}) \propto L(\vec{\theta})^{1/T} p(\vec{\theta}),$$

where $L$ and $p$ are respectively the likelihood and prior distributions.

For high $T$, individual peaks in $L$ become flatter and broader, making the distribution easier to sample via MCMC. A set of $N$ chains is assigned temperatures in a ladder $T_1 < T_2 < \ldots < T_N$, with $T_1 = 1$ (the target temperature). The temperatures are typically geometrically spaced from 1 up to some $T_{\text{max}}$, decided in advance (a convention that we shall discuss in more detail in §2.2).

Each chain is allowed to explore its tempered distribution $\pi_T$ under an MCMC algorithm, while at predetermined intervals “swaps” are proposed between (usually adjacent$^1$) pairs of chains and accepted with probability

$$A_{i,j} = \min \left\{ \left( \frac{L(\vec{\theta}_i)}{L(\vec{\theta}_j)} \right)^{\beta_j - \beta_i}, 1 \right\},$$

where $\vec{\theta}_i$ is the current position in the parameter space of the $i$th chain and $\beta_i \equiv 1/T_i$ is the inverse temperature of this chain. When a swap is accepted, the chains exchange their positions in the parameter space, so that chain $i$ is at $\vec{\theta}_j$ and chain $j$ is at $\vec{\theta}_i$. Since the hottest chains can access all of the modes of $\pi$ (as long as $T_{\text{max}}$ is chosen appropriately), their locations propagate to colder chains, ultimately allowing the $T = 1$ (cold) chain to efficiently explore the entire target distribution. At the same time, the positions of the colder chains propagate upward to higher temperature chains, where they are free to explore the entire prior volume.

The goal in choosing an effective ladder of temperatures is to minimise the ACT of the cold chain (our measure of the efficiency of the sampler). The requirements to this end are two-fold:

(i) $T_{\text{max}}$ must be large enough that isolated modes of $L$ broaden sufficiently that an individual MCMC chain can efficiently access all of these modes when sampling under the tempered posterior $\pi_T$ in §1 at $T = T_{\text{max}}$. We denote this temperature $T_{\text{prior}}$.

(ii) Since $A_{i,j}$ depends on $\beta_i - \beta_j$, the differences between temperatures must be small enough that neighbouring chains can communicate their positions efficiently with one another.

Both requirements depend sensitively on the (unknown) shape of the target distribution, so it is difficult to select temperatures appropriately in advance.

In choosing $T_{\text{max}}$, one must know roughly the relative size and separation of the modes to be explored. As an example, consider a one-dimensional likelihood with two Gaussian modes of width $\sigma = 1$ and centres $\mu = 1 \pm 10$. In order to prevent a sampler from getting stuck on one of the modes, they must be widened to roughly the separation between them$^2$, giving $\sigma = O(10)$. The width of a Gaussian peak scales with the temperature as $\sqrt{T}$, so we might choose $T_{\text{max}} = 100$; Figure 1 illustrates the resulting coalescence of the modes. A different configuration of modes will, of course, require a different $T_{\text{max}}$.

On the other hand, the swap acceptance probability $A_{i,j}$ depends on the distribution of likelihood values at temperatures $T_i$ and $T_j$. In the case of a likelihood distribution comprising a single Gaussian mode, the time-averaged acceptance ratios between chains, $E[A_{i,j}]$, can be computed analytically (see §2.2).

In general, we don’t know in advance what the target distribution looks like, and so choosing an effective ladder becomes a heuristic exercise, relying largely on educated guesswork. We are therefore motivated to find some method of empirically determining an effective ladder.

2.1 Ladder selection

For an $n$-dimensional problem, the conventional choice of temperatures is a geometrically spaced ladder constructed so that approximately 23% of swaps proposed between chains will be accepted when sampling from an $n$-dimensional, unbounded Gaussian distribution (Earl & Deem 2005; ter Braak & Vrugt 2008; Roberts & Rosen-
A consequence of this strategy is that increasing the number of chains $N$ does not improve communication between existing chains, which is determined by $E[A_{i, j}] = 0.23$. Instead, adding new chains extends the ladder to higher temperatures. This may be appropriate for an unbounded posterior, but for a realistic problem with a finite prior volume, the acceptance ratio between adjacent chains saturates to $\sim 100\%$ at some temperature $T_{\text{prior}}$, at which the posterior $\pi_T$ begins to look like the prior $p$.

For this geometric spacing scheme – where $T_{\text{prior}}$ is unknown – there is therefore an optimal number of chains, $N_{\text{opt}}$, such that $T_{\text{prior}} \approx T_{N_{\text{opt}}} \cong T_{\text{max}}$. For $N < N_{\text{opt}}$ none of the chains will be sampling from the prior (so the sampler may not find all of the modes), while for $N > N_{\text{opt}}$ we end up with several chains sampling redundantly from the prior.

Since we are generally ignorant of $T_{\text{prior}}$ for the problem at hand, we are motivated to find an alternative temperature selection strategy.

It has been suggested in the literature (Earl & Deem 2005, Sugita & Okamoto 1999, Kofke 2002, 2004) that one could select temperatures such that the acceptance ratios $A_{i, j}$ are uniform for all pairs $(i, j)$ of adjacent chains, in an attempt to ensure that each sample sequence $\theta(t)$ for $t = 1, 2, \ldots$, as it moves between chains, spends an equal amount of time at every temperature. Sugita & Okamoto (1999) justify this notion experimentally – in the context of molecular dynamics – with test cases in which such a ladder indeed performs well. They use an algorithm derived from that of Hukushima & Nemoto (1996), which selects temperatures according to an iterative process for which a uniform-$A$ ladder is a fixed point. Earl & Deem (2005) provide further references for similar methods of determining temperature ladders that yield a given a target acceptance ratio (Rathore et al. 2005, Sanbonmatsu & García 2002, Schug et al. 2004). However, these methods do not address requirement (i) discussed above, that the temperature ladder should reach a $T_{\text{max}}$ sufficient for all of the modes of $L$ to mix (specified by $T_{\text{prior}}$).

Kofke (2002) discusses the selection of temperature ladders in the context of molecular simulations. He shows that, in simulations of such thermodynamic systems, there is a close relation between the specific heat of the system, $C_V$, and the acceptance ratios between adjacent temperatures. In particular, when $C_V$ is constant with respect to $T$ over a given temperature interval, then a geometric spacing of temperatures on that interval yields uniform acceptance ratios between adjacent temperatures.

In the language of thermodynamics, the energy of the system, $U$, is analogous to $-\log L$, and an analogue to the specific heat can therefore be defined as

$$C_V(T) = -\frac{d}{dT} E[\log L]_T,$$

where $E[\cdot]_T$ denotes the expectation operator over $\theta$ under the distribution $\pi_T(\theta)$. $E[\log L]_T$ is therefore the expectation of the untempered log likelihood collected when sampling from the posterior at temperature $T$.

In the context of Bayesian inference, Kofke’s result therefore tells us that if the mean log likelihood collected by a sampler responds linearly to changes in temperature, then a geometrically spaced temperature ladder will achieve uniform acceptance ratios between adjacent chains. Conversely, temperature intervals on which $E[\log L]_T$ is strongly non-linear in $T$ represent a phase transition that will require more careful placement of temperatures, as we shall show in §2.4.5.

### 2.2 The ideal Gaussian distribution: a simple example

In the simple case of a unimodal Gaussian likelihood under a flat prior, the optimal temperature spacing at low temperatures – where very little likelihood mass is truncated by the prior – can be analysed by approximating the prior to be unbounded. We show that, for this tractable example, a geometric temperature spacing is consistent with both the uniform-$A$ criterion and also with the alternative criterion that the Kullback–Leibler (KL) divergence is uniform between all pairs of adjacent chains. We use the example to illustrate the relationship between the analytical distribution of $\log L$, the acceptance ratio $A_{i, j}$, and the temperature $T$.

We shall work with an $n$-dimensional unit Gaussian centred on the origin (the same result can be achieved for a general Gaussian through a simple change of coordinates). Since the prior is uniform and unbounded, we can restrict attention to the likelihood distribution $L$.

---

3 The approximation breaks down at higher temperatures, where boundary effects become significant. Indeed, with no prior boundaries, there is no $T_{\text{prior}}$ at which the mode is spread over the entire prior volume.
In this case, the probability density \( \tilde{p} \) for the values of \( \log L(\bar{\theta}) \) collected by the sampler is

\[
\tilde{p}(\log L) = \frac{\exp(\log L - \log L_0)}{\Gamma(\frac{2}{\beta})}, \tag{4}
\]

where \( L \) is normalised so that \( \log L(\bar{\theta}) = 0 \) and \( n \) is the number of parameters.

At a temperature \( T \), \( -\log L \) simply follows a gamma distribution \( \Gamma(\alpha, \beta) \) with shape parameter \( \alpha = n/2 \) and rate parameter \( \beta = 1/T \). Thus, for a chain sampling at temperature \( T \), the log likelihood distribution is \( \tilde{p}(\log L) = T \tilde{p}(\log L/T) \).

Over long time-scales, the average acceptance ratio between chains \( i \) and \( j \) is

\[
E[A_{i,j}] = \int_{-\infty}^{\infty} A_{i,j} \tilde{p}_T(\log L_i) \tilde{p}_T(\log L_j) \, d\log L_i \, d\log L_j
\]

\[
= \left( \frac{1}{\sqrt{\beta}} 2^{n-1} \gamma_{i,j}^{n/2} \Gamma \left( \frac{n+1}{2} \right) \right) \cdot \left( 2 \tilde{F}_1 \left( \frac{n}{2}, \frac{n}{2}; \frac{n}{2} + 1; -\frac{1}{\gamma_{i,j}} \right) - \gamma_{i,j}^{n/2} 2 \tilde{F}_1 \left( \frac{n}{2}, \frac{n}{2}; \frac{n}{2} + 1; -\gamma_{i,j} \right) \right) + 1, \tag{5}
\]

where \( \tilde{F}_1 \) is the regularised Gauss hypergeometric function and \( \gamma_{i,j} = T_j/T_i \) is the ratio between the temperatures of two chains. Since \( E[A_{i,j}] \) depends on \( T_i \) and \( T_j \) only through the ratio \( \gamma_{i,j} \), uniform acceptance ratios between all adjacent pairs of chains can be achieved with a geometric spacing of temperatures - where \( \gamma_{i,i+1} \) is constant - for a unimodal Gaussian likelihood.

The log spacing required for a particular acceptance ratio also depends on the dimension of the parameter space, with more parameters requiring a closer spacing of temperatures, illustrated by Figure 3. This can be understood by looking at the expectation and variance of \( \log L \) at a particular temperature (see Figure 2).

\[
E[\log L]|_T = -\frac{nT}{2} \quad \text{and} \quad \text{Var}[\log L]|_T = \frac{nT^2}{2}. \tag{6}
\]

Note that the specific heat from \([3]\) is a constant \( n/2 \), as expected.

Since the acceptance ratio \( A_{i,j} \) depends on \( \log L_i - \log L_j \), the more separate the distributions of \( \log L_i \) and \( \log L_j \) at their respective temperatures, \( T_i \) and \( T_j \), the lower the acceptance ratio between such chains will be. For two chains at temperatures \( T \) and \( \gamma T \), the separation of the means of \( \tilde{p}_T \) and \( p_T \), in units of the standard deviation at \( T \), will be

\[
\frac{E[\log L]|_T - E[\log L]|_{\gamma T}}{\sqrt{\text{Var}[\log L]|_T}} = (\gamma - 1) \sqrt{\frac{n}{2}}. \tag{7}
\]

It follows that – for constant \( \gamma \) as the dimension \( n \) increases, so the acceptance ratio between chains at temperatures \( T \) and \( \gamma T \) falls. For a higher dimensional target distribution, therefore, a closer spacing of temperatures is required for a given acceptance ratio.

For more general distributions, by considering the overlap of \( \tilde{p}_T(\log L) \) at different temperatures, Falcioni & Deem (1999) argue that the number of temperatures \( N \) required to efficiently sample the posterior distribution should scale with \( \Delta \log L/\sqrt{n} \), where \( \Delta \log L \) is the range of \( E[\log L]|_T \) between \( T = 1 \) and \( T = T_{\text{prior}} \). That is:

\[
N \propto \frac{E|\log L_1| - E[\log L]_{T_{\text{prior}}}}{\sqrt{n}}. \tag{8}
\]

Since the log likelihood range \( \Delta \log L \) itself depends on the dimension of the system \( n \), it is difficult to apply this relation in practice. However, for the ideal Gaussian, we can see from \([4]\) that \( \Delta \log L \) scales with \( n \), and so \( N \) scales with \( \sqrt{n} \), as we might expect.
2.3 The Kullback–Leibler divergence

Another measure of the optimal spacing of temperatures is the Kullback–Leibler (KL) divergence between adjacent chains. The KL divergence from a hot distribution \( \pi_{T_j} \) to a cold distribution \( \pi_{T_i} \),
\[
D_{KL}(\pi_{T_i} \| \pi_{T_j}) = \int \pi_{T_i}(\theta) \log \frac{\pi_{T_i}(\theta)}{\pi_{T_j}(\theta)} \, d\theta,
\]
quantifies the information gained about the posterior with each step down the temperature ladder, from the prior \( \pi_T \equiv \pi \) to the posterior \( \pi \equiv \pi_{T \to \infty} \). It is reasonable to expect that for an optimally-spaced ladder – that is, one with a minimal ACT on the cold chain for a given number of chains – the information gain should be uniform for every step down the ladder.

For the example of the ideal Gaussian of \( \overline{2.2} \) the KL divergence is, straightforwardly,
\[
D_{KL}(\pi_{T_i} \| \pi_{T_j}) = \frac{n}{2} \left( \frac{1}{\gamma_{i,j}} + \log \gamma_{i,j} - 1 \right).
\]

Like the swap acceptance ratio, therefore, uniform KL divergence over the entire ladder is also achieved by a geometric spacing of temperatures for the ideal Gaussian.

Unfortunately, unlike the acceptance ratio, the KL divergence is difficult to compute numerically while sampling, owing to the unknown – and temperature-dependent – evidence (normalisation) values on \( \pi_{T_i} \) and \( \pi_{T_j} \).

We henceforth assume that spacing temperatures for uniform acceptance ratios is a reasonable approximation of a ladder that is optimal in the ACT of the cold chain. We make this assumption on faith and, while we present a brief examination of its validity in \( \overline{4.1} \) it invites a more careful study.

3 ADAPTIVE TEMPERATURE LADDERS

From the arguments in \( \overline{2.2} \) and the references therein, we shall assume that uniformity of acceptance ratios provides a good approximation to the optimal temperature ladder for parallel tempering problems. In this section, we describe an algorithm for dynamically adapting chain temperatures to achieve uniform acceptance ratios for inter-chain swaps.

From \( \overline{2.2} \), as \( 1/T_j - 1/T_i \to 0 \), \( A_{i,j} \to 1 \), so in order to increase the expected acceptance ratio between chains, it suffices to move them closer together in temperature space; conversely, to reduce \( \mathbb{E}[A_{i,j}] \), we can push the chains apart. We will henceforth adopt the notation that \( A_i \equiv A_{i,i-1} \) and that \( T_i < T_{i+1} \), with \( T_1 = 1 \) being the untempered or cold chain (which samples from the target distribution, \( \pi \)). Here, \( A_i(t) \) are the instantaneous acceptance ratios between chains, but we shall shortly describe the discrete case where empirical measurements of \( A_i \) are collected with each iteration of the sampler.

3.1 Dynamics

Our goal is to dynamically adjust the temperatures of the chains to achieve uniform acceptance ratios as we sample the target distribution. We define our temperature dynamics in terms of the log of the temperature difference between chains,
\[
S_i \equiv \log(T_i - T_{i-1}).
\]

Under this scheme, finite changes to \( S_i \) will always preserve the correct ordering of temperatures (\( T_1 < ... < T_N \)).

To achieve the same \( A_i \) for all chains, we can drive the gap \( S_i \) according to the acceptance ratios between chain \( i \) and those immediately above and below, to wit
\[
\frac{dS_i}{dt} = \kappa(t) \left[ A_i(t) - A_{i+1}(t) \right],
\]
for \( 1 < i < N \), where \( \kappa \) is a positive constant controlling the time-scale of the evolution of \( T_i \). \( \kappa \) can be interpreted as the instantaneous exponential time-constant for temperature adjustments. The two extremal temperatures, \( T_1 \) and \( T_N \), are fixed (see below).

Under this scheme, chain \( i \) will attempt to increase the gap in temperature space between itself and chain \( (i+1) \) if swaps are accepted too often and close it when they are accepted too seldom — and similarly for chain \( (i-1) \) — equilibrating at \( A_i \) that are uniform over \( i \). Therefore, for an appropriate choice of \( \kappa \) – discussed momentarily – these rules drive the chains \( i = \{2, \ldots, N-1\} \) toward even acceptance spacing.

However, in order to efficiently sample a target distribution with strongly separated modes (such that a traditional MCMC sampler would be unable to traverse the “valleys” between them), \( T_N \) must be high enough that the modes are flattened out and the chain can explore the entire parameter space unhindered. This amounts to the topmost chain sampling from the prior distribution\(^4\), which we achieve trivially by setting the inverse temperature of this chain as \( \beta_0 = 0 \).

This continuous system is discretised as
\[
S_i(t + 1) - S_i(t) = \kappa(t) \left[ A_i(t) - A_{i+1}(t) \right],
\]
where \( A_i(t) \) are the acceptance ratios accumulated by the sampler at the current iteration.

The values of \( A_i \) are measured empirically at each iteration as the fraction of swap proposals between chains that were accepted. For a traditional sampler comprising one sample per chain, these will be either 0 or 1. For ensemble samplers, however, comprising many distinct walkers per temperature, there is a finer granularity in the measurements of \( A_i \).

Importantly, the temperature adjustment scheme we have proposed – and, more generally, any adaptive sampling scheme – in fact violates the condition for detailed balance that ensures that an MCMC sampler will converge to the target distribution. Roberts & Rosenthal (2007) investigate the conditions required of such an adaptive sampler for it to be ergodic in the target distribution – that is, that it will converge on long time-scales. They determine (from their Theorem 1 and Corollary 4) that diminishing the amplitude of adaptations in the

\(^4\) For analytic priors, this special case, where the likelihood is ignored, can be treated separately by having the sampler draw independent samples directly from the prior.
transition kernel with each iteration is sufficient for the sampler to be ergodic in the target distribution. We therefore suppress temperature adjustments to ensure that the sampler is Markovian on sufficiently long time-scales\(^5\).

The rate of diminution of temperature adjustments is a trade-off between the rate of convergence of the temperature ladder and that of the sampler itself toward its stationary distribution. We modulate the dynamics with hyperbolic decay to suppress the dynamics on long timescales,

\[ \kappa(t) = \frac{1}{\nu t + t_0}, \quad (14) \]

where \( t_0 \) is the time at which the temperature adjustments have been reduced to half their initial amplitude. The initial amplitude of adjustments is in turn set by \( \nu \), the time-scale on which the temperatures evolve.

### 3.2 Parameter choice

In the scheme of \((13)\) and \((14)\), there are two parameters to choose: \( t_0 \) and \( \nu \). The dynamical time parameter \( t \) in \((13)\) is measured in units of intra-chain jumps of the sampler, with temperature adjustments being made at every iteration.

The lag parameter \( t_0 \) sets the time-scale for the attenuation of temperature adjustments. This decay factor in \( \kappa \) is included as a fail-safe mechanism to ensure that, even for target distributions on which the temperature dynamics fail to find an equilibrium set of temperatures, the ladder will always converge over long time-scales. This condition guarantees that the sampler correctly explores the target distribution.

\( t_0 \) should therefore be chosen to set an upper limit on the time-scale of convergence to catch such badly behaved target distributions. For our tests, we use \( t_0 = 10^3 \), though other applications might require different values.

Meanwhile, \( \nu \) sets the overall time-scale of temperature adjustments and is used to smooth out statistical errors on the measurements of the acceptance ratios \( A_i \). In general, for a sampler of \( n_w \) walkers, the acceptance count per iteration, \( n_w A_i \), is a random variable that follows a binomial distribution \( B(n_w, E[A_i]) \), so that \( A_i \) has variance

\[ \text{Var}[A_i] = \frac{E[A_i](1 - E[A_i])}{n_w}. \]

Since the dynamical equations \((13)\) are linear in \( A_i \), they will be driven by the means, \( E[A_i] \), on long time-scales, assuming that the noise in the system from counting errors – proportional to \( 1/\sqrt{n_w} \) – does not cause short-term changes in \( E[A_i] \).

We should therefore choose \( \nu \) to be large enough that, even with few walkers, a sampler is not susceptible to large statistical fluctuations in its measurements of \( A_i \) while, conversely, it is not too slow to equilibrate.

A good choice of \( \nu \) depends on the response of \( E[A_i] \) to changes in the relevant chains’ temperatures, and therefore depends on the particular likelihood function that is being sampled. However, if \( E[A_i] \) will eventually be of order, say, 0.25, and we want the measurements of \( A_i \) to be between 0.2 and 0.3, then we should average \( A_i \) over 100 swap proposals. In practice, we have found that \( \nu = 100 \) indeed works well on our test cases.

### 4 EXAMPLES

We have implemented the algorithm proposed above as a modification to the ensemble sampler emcee of Foreman-Mackey et al. (2013). Our implementation can be found at https://github.com/willvousden/emcee/tree/adaptive.

In this section we apply our implementation to specific examples in order to understand how and when the traditional geometric spacing fails and how much the uniform-A strategy might help us. We present the following test cases.

(i) In §4.1 we compare the uniform-A strategy used by the temperature dynamics of \((13)\) with the alternative strategy of uniform KL divergence discussed in §4.2 on the example of a unimodal truncated Gaussian likelihood.

(ii) In §4.2 we test the dynamics on a more complex, bimodal distribution for various choices of the number of chains \( N \). We compare the resulting ACTs of the sampler with those of another sampler using a geometric ladder whose maximum temperature is fixed such that \( T_{\max} \approx T_{\text{prior}} \).

(iii) In §4.3 we test the algorithm against the more difficult egg-box distribution with 243 modes. For comparison, we sample from the same distribution with a geometric ladder constructed to yield 25% acceptance ratios when applied to the ideal Gaussian discussed in §4.2.

For all of these tests, \( \nu = 10^2 \) and \( t_0 = 10^3 \) are used to control the dynamics in \((14)\).

#### 4.1 Truncated Gaussian

Our first test case is an \( n \)-dimensional, unimodal, unit Gaussian similar to that of \((2.2)\) but with finite prior volume. The simplicity of this case admits some exact analysis before recourse to numerics, which allows us to test the approximations made in \((2.2)\).

At low temperatures, where the prior boundaries do not truncate much of the likelihood probability mass, the optimal temperature spacing should be similar to that of the ideal Gaussian. By imposing a step-like cut-off in the prior at a radius of \( R \), there will be some temperature at which this approximation will fail and a geometric spacing becomes inappropriate.

For the likelihood we use the same distribution as in \((2.2)\) while for the prior we use a uniform distribution over the closed \( n \)-ball of radius \( R = 30 \), centred on the origin. The likelihood and prior are defined by

\[ L(\vec{\theta}) \propto \exp \left( -\frac{1}{2} \| \vec{\theta} \|^2 \right), \quad (15) \]

\[ p(\vec{\theta}) \propto \begin{cases} 1 & \text{if } \| \vec{\theta} \| \leq R, \\ 0 & \text{otherwise,} \end{cases} \quad (16) \]

\( \odot 2002 \text{RAS, MNRAS 000} \).
where $\| \cdot \|$ is the Euclidean norm on $\mathbb{R}^a$. Subsequently, the normalised posterior generated by (15) and (16) at temperature $T$ is

$$\pi_T(\theta) = \left\{ \begin{array}{ll} \frac{(2\pi)^{\frac{a}{2}} r(\frac{a}{2})}{\tilde{\gamma}(\frac{a}{2}, \frac{R^2}{2T})} \exp\left(-\frac{\|\theta\|^2}{2T}\right) & \text{if } \|\theta\| \leq R, \\ 0 & \text{otherwise}, \end{array} \right. $$

where $\tilde{\gamma}(a, z)$ is the lower incomplete gamma function.

In the low-temperature limit, this distribution converges to ideal Gaussian distribution. We should therefore expect the KL divergence for a step down the temperature ladder to asymptote to (10) as $T \to 0$, where the effects of the prior boundary are negligible. Indeed, the KL divergence of (17) from $T_2$ to $T_1$ is available analytically as

$$D_{KL} = -\frac{(T_2 - T_1) \tilde{\gamma} \left(1 + \frac{a}{2}, \frac{R^2}{2T} \right)}{T_2 \tilde{\gamma} \left(\frac{a}{2}, \frac{R^2}{2T_2} \right)} + \frac{n}{2} \log \left(\frac{T_2}{T_1}\right) + \log \left(\frac{\tilde{\gamma}(\frac{a}{2}, \frac{R^2}{2T_2})}{\tilde{\gamma}(\frac{a}{2}, \frac{R^2}{2T_1})}\right).$$

If we set $T_2 = \gamma T_1$ (with $\gamma T_1 \ll 1$), then $\tilde{\gamma}(a, z) \to \Gamma(a)$ as $T_1 \to 0$, and the expression reduces to (10), as expected.

Figure 4 illustrates this convergence for $n = 5$. The point on this plot at which the solid line diverges from the dashed line, for each $\gamma$, predicts the temperature beyond which a geometric spacing of temperatures is no longer optimal (for optimality as defined by uniform KL divergence between chains). This is caused by truncation of the tempered likelihood by the prior boundaries.

Of course, since the KL divergence cannot easily be assessed empirically by an MCMC sampler, and we must instead resort to using acceptance ratios, we would like to know how consistent these two schemes are outside the assumptions of (2.2).

Figure 5 shows contours of constant $D_{KL}$, calculated from (18), and contours of constant $A_1$, illustrated by points representing temperature pairs (from ladders selected by the algorithm developed in [13]). In the low temperature limit, as expected, both schemes select a geometric spacing of temperatures consistent with the ideal Gaussian of (2.2) (i.e. the contours remain constant in $\gamma$). At higher temperatures, both schemes depart from the geometric spacing, but they do so differently. The uniform acceptance scheme displays a more gradual departure from a geometric spacing than the contours of constant $D_{KL}$. The smaller $\gamma$ selected by the uniform-A scheme outside the geometric regime, however, suggest that closer spacing is required in difficult temperature ranges (e.g., across a phase transition) in order to achieve uniform $A$ than would be required for uniform $D_{KL}$. There is therefore less risk of a large gap in temperature across such a temperature range, at the cost of (potentially) slightly less efficient communication across the rest of the ladder. The uniform-A criterion for optimality is therefore conservative with respect to a uniform-$D_{KL}$ criterion.

We can also visualise the ladder specification in terms of the density of chains over temperature. We define this density, in log $T$, as

$$\eta(\log T) = \frac{dN}{d\log T} = \frac{1}{\Delta \log T} = \frac{1}{\log \gamma},$$

with $\gamma = T_{i+1}/T_i$, where $T_{i+1}$ and $T_i$ are the chain temperatures to either side of $T$.

Figure 6 shows this density for a temperature ladder of 20 chains that is in equilibrium under the temperature dynamics of $\gamma = 2$ (the $N = 20$ contour of Figure 5). The density exhibits the expected uniformity of $\gamma$ for low temperatures but falls for $T \gtrsim 80$. The width $\sigma$ of the unit Gaussian at temperature $T = \sqrt{T}$, so at this temperature the prior boundary is at $\sim 3\sigma$. At $T = 80$, $\sim 5\%$ of the likelihood mass is truncated -- compared to $< 0.1\%$ for $T = 40$ and $\sim 35\%$ for $T = 160$ -- indicating that the prior boundary becomes significant in this temperature regime.

This drop in density reflects the convergence of the tempered posterior distribution, $\pi_T$, toward the prior as $T \to \infty$. As $\pi_T$ becomes flatter, fewer chains are needed per log $T$ to maintain good communication.

Also shown on figure Figure 6 is the square root of the estimated specific heat $C_V$ of the system as discussed in [2.1] which can be seen to track closely the logarithmic chain density $\eta$ when appropriately normalised. While the provenance of this relationship is unclear, it demonstrates the relevance of the specific heat in determining an effective temperature ladder.

---

6 While we do not consider $T < 1$ in our simulations, the case of $T \to 0$ can equivalently be thought of as $R \to \infty$, since the width of the Gaussian scales with $\sqrt{T}$. 
Figure 5. A contour plot of the KL divergence, or information gain, from a hot chain at temperature $T_{\text{high}} = \gamma T_{\text{low}}$ to a colder chain at temperature $T_{\text{low}}$, both sampling from the Gaussian likelihood (17). The coloured lines show the equilibrium $N$-chain temperature ladders reached by the temperature dynamics algorithm of §3, where the acceptance ratio is the same between any pair of adjacent chains. The points on these lines represent pairs of adjacent temperatures $(T, \gamma T)$ (excluding the top-most, where $\gamma = \infty$).

Figure 6. Orange: The density of chains per log $T$ under the truncated Gaussian distribution (17), where $N = 20$, $n = 25$, and temperatures are chosen for uniform acceptance ratios between chains. The chains have equilibrated to 77% acceptance. Blue: The square root of the specific heat of the truncated Gaussian distribution, normalised to match the chain density $\eta$ of the uniform-$A$ ladder, between $T_1$ and $T_{N-1}$. The specific heat $C_V$, from (3), is estimated from the sample means of log $L$ over several runs using different temperature ladders.

4.2 Double Rosenbrock function

The previous test demonstrated how a geometric ladder spaces temperatures too closely at higher temperatures, as the prior boundary becomes significant. While this may be an inefficient use of resources, it at least doesn’t drastically inhibit communication between high temperatures and low temperatures. Instead, we now turn to a more complex, bimodal likelihood distribution for which a geometric spacing might cause bottlenecks in the communication between high and low temperatures.

We use a likelihood derived from the two-dimensional Rosenbrock function $f$:

$$L(x, y) \propto \frac{1}{c + f(x, y)} + \frac{1}{c + f(-x, y)} T_p,$$

where $f(x, y) = (a - x)^2 + b(y - x^2)^2$. (21)

$T_p$ is a pre-tempering factor chosen to increase the contrast of the distribution, making it harder to sample. When $T_p \ll 1$, each mode is locally Gaussian, making the results comparable to the Gaussian example considered in §2.2.

For the following tests, we use $a = 4$, $b = 1$, $c = 0.1$, and $T_p = 10^{-3}$. We use a flat prior on $[-10, 10] \times [-20, 100]$. Figure 7 illustrates this likelihood over the prior volume.

4.2.1 Test: temperature evolution

As an illustrative example, we first tested the temperature dynamics of §3 with the double Rosenbrock posterior distribution (20) using 13 chains. Figure 8 shows the evolution of the temperature ladder according to these dynamics, while Figure 9 shows the chain density $\eta(\log T)$ for the equilibrated temperature ladder.
Temperature dynamics for PTMCMC samplers

Figure 8. The evolution of ladder of 13 temperatures $T_i$ and acceptance ratios $A_i$ over an emcee run of $10^6$ iterations under the Rosenbrock likelihood. Chains 1 and 13 are not shown, having fixed temperatures $T_1 = 1$ and $T_{13} = \infty$.

While the equilibrated chains are distributed uniformly in $\log T$ for $T \lesssim 50$, there is a distinct peak in $\eta$ at $T \approx 800$, where a simple geometric spacing of temperatures hinders communication between chains. This peak occurs at a phase transition where the two modes of the likelihood distribution begin to mix and $E[\log L]$ changes rapidly with $T$, indicated by the sharp change in specific heat in the bottom panel of Figure 9. Since the shape of the likelihood distribution in this regime becomes very sensitive to $T$, a higher density of chains is needed to maintain a given acceptance ratio. We also note that in the geometric regime (i.e. for low $T$) the specific heat is approximately $n/2 = 1$, with $E[A] \approx 57\%$, consistent with the values derived for the ideal Gaussian from and respectively.

Finally, however, the figure of merit for a temperature specification in a PTMCMC simulation is the resulting ACT for the target temperature ($T = 1$) of the sampler. We must therefore test the performance of the sampler empirically.

We use the term ACT to refer to the integrated autocorrelation time discussed by Sokal (1997), which we estimate according to the algorithm used in the acor package (see Appendix A and http://www.math.nyu.edu/faculty/goodman/software/acor/ for details).
following tests, we use the ACT of the first parameter, \(x\), as a measure of the efficiency of the sampler (since \(\frac{\tau}{\tau_{\text{geo}}} \approx 4\) is bimodal in \(x\) but unimodal in \(y\)).

### 4.2.2 Test: improvement over a geometric ladder

In [2], we claimed that aiming for uniform acceptance ratios between chains yields a good temperature ladder. Specifically, we expect that a ladder selected for uniform acceptance ratios should lead to a lower ACT for the \(T = 1\) chain than that resulting from a plain geometric ladder.

The geometric ansatz that we use has a fixed maximum temperature such that \(T_N = 2 \times 10^4\). As \(N\) increases, more chains are added between \(T_i\) and \(T_N\), maintaining the geometric spacing. Under this arrangement, the addition of new temperatures is not redundant even when \(T_N\) is already high enough to sample from the prior; the additional chains instead aid inter-chain communication at lower temperatures. Since \(T_N\) is close to the temperature at which the posterior becomes the prior, there is little CPU time wasted in sampling redundantly from the prior with several chains, while lower-temperature chains can still communicate with a chain sampling from the prior. Under this set-up, therefore, the ACT always decreases as \(N\) increases, per Figure 10.

To test the improvement in ACT, \(\tau\), conferred by our temperature dynamics, we allowed emcee to explore the target distribution \((20)\) with different numbers of chains, \(N\), using both the uniform-\(A\) ladders and geometrically spaced ladders. The resulting ACTs, \(\tau_{\text{geo}}\) and \(\tau_{\text{acc}}\), are plotted against \(N\) in Figure 10.

In this example, an \(N\)-chain ladder dynamically adapted for uniform acceptance ratios clearly outperforms a geometrically spaced ladder of the same size for all \(N\).

The benefit of a uniform-\(A\) ladder is most pronounced at low \(N\) – i.e., where there are few chains available. In this regime, the sampler will be more sensitive to phase transitions, since the bigger gaps in temperature could cause severe bottlenecks in communication across the temperature ladder.

When \(N\) is large, the differences in acceptance ratios between a geometric ladder and one chosen for uniform \(A\) becomes less significant. In this case, the difference between the limiting (minimum) acceptance ratio for a ladder and the ladder’s average acceptance ratio is proportionally smaller.

In the case of the double Rosenbrock distribution \((20)\), we have found that, once the minimum acceptance ratio for a geometric ladder (terminating at \(T_{\text{max}} = 2 \times 10^4\)) exceeds \(\sim 10\%\), reallocating temperatures for uniform acceptance ratios does not reduce the measured ACT by more than 25\%. This occurs when \(N \approx 7\) in the current example. Nonetheless, there remains an overall improvement in ACT regardless of \(N\).

Figure 10 also shows, in the middle pane, the total number of iterations per independent sample across all chains. This quantity, given by \(N \times \tau\), is proportional to the total CPU time of the simulation, while \(\tau\) itself is proportional to the CPU time per chain, or wall time, of the simulation. In this instance, the CPU time of a run diminishes with \(N\) in much the same fashion as the wall time does. The fractional improvement in CPU time is of course the same as for wall time – \(\tau_{\text{geo}} / \tau_{\text{acc}}\).

### 4.2.3 Test: chain removal

To determine whether a uniform-\(A\) temperature placement strategy is in fact close to optimal, we assess the contribution of each chain from such a temperature ladder to the efficiency of the sampler, as measured by its ACT. If this contribution is equal for all chains, then we can conclude that it is indeed optimal to have them all exchanging equally – that is, with uniform acceptance ratios.

To this end, we conducted the following test:

(i) Sample from (20) with \(N = 7\) chains under the temperature dynamics of \([3]\) until the temperatures have equilibrated to \((T_1, \ldots, T_7)\) to give uniform acceptance ratios.

(ii) Generate 5 new test ladders, each of 6 chains, formed by removing the \(i\)th chain from that determined above – i.e. \((T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_7)\) – for \(i = 2, \ldots, 6\).
Autocorrelation

\[ \tau \]

Figure 11. The cold-chain ACTs for samplers exploring the double Rosenbrock distribution (20) per the test described in §4.2.3. The points denote the ACTs from ladders generated according to the scheme in §4.2.3. The dashed and dotted lines represent for ladders of the same \( N \). The approximate ACTs are 504, 531, 551, 736, and 765, for the solid lines correspond to the ladders generated by the sampler exploring the double Rosenbrock distribution (20). The cold-chain autocorrelation function for a ladder constructed to give a fixed acceptance ratio of \( \tau_{\text{acc}} \) < \( \tau_{\text{test}} \) < \( \tau_{\text{geo}} \) for ladders of the same \( N \), \( \tau_{\text{test}} \) increases with the temperature of the chain that is removed, suggesting that additional chains are more useful at higher temperatures.

The sharp jump in \( \tau_{\text{test}} \) when a chain above \( T \approx 200 \) is removed arises from the phase transition that occurs as \( T \) approaches \( T_{\text{prior}} \), indicated by a peak in \( C_V \) (visible in Figure 9).

We can understand this behaviour by examining the complete autocorrelation functions from which these ACTs are estimated. Illustrated in Figure 12 these autocorrelation functions exhibit two distinct time-scales. Firstly, there is a large autocorrelation for lags \( \lesssim 100 \) for all \( i \) – particularly \( i = 2 \) – corresponding to the ACT of the sampler within one of the two modes: that is, the time taken for the sampler to generate an independent sample without changing mode. Secondly, there is a visible hump in the autocorrelation function for 100 \( \lesssim \) lag \( \lesssim 2000 \), corresponding to the time taken for the sampler to migrate between modes. Removing the second chain from initial geometric ladder of 7 chains increases the intra-mode ACT in particular, but does not affect the inter-mode ACT. Meanwhile, while removing higher temperature chains pushes the secondary hump outward to larger lags, increasing the inter-mode ACT instead.

The overall autocorrelation time in which we are interested, discussed by Sokal (1997) and in Appendix A represents the time between independent samples of the system. It is therefore set by the time-scale on which the sampler migrates to a new mode independently of the current mode. Removing a chain at higher temperatures increases the inter-modal ACT, and therefore damages the efficiency of the sampler.

Nonetheless, all of the tested temperature ladders yielded lower ACT than the default geometric ladder, despite the geometric ladder being chosen with prior knowledge of \( T_{\text{prior}} \).

4.3 Egg-box in five dimensions

To test the algorithm’s performance on a yet more strongly multi-modal distribution, we use an egg-box distribution defined by the likelihood

\[
L(\hat{\theta}) \propto \left( \frac{1}{2} \prod_{i=1}^{n} \cos \theta_i + \frac{1}{2} \right)^{T_p}.
\]

(22)

For a small value of the pre-tempering factor \( T_p \) the modes of this distribution become locally Gaussian, and in the low-\( T \) regime should therefore generate results similar to those of the Gaussian distributions examined in §2.2 and §4.1. For the following tests, we choose \( T_p = 10^{-3} \).

We explore this likelihood distribution in 5 dimensions over a flat prior on \([-L/2, L/2]^n\), where we choose \( L = 3\pi \), giving \( 3^n = 243 \) modes.

Rather than compare our uniform-\( A \) temperature ladder against a geometric ladder with a fixed maximum temperature, as in §4.2, we instead use a geometric ladder constructed to give a fixed acceptance ratio of \( \text{E}[A] = 0.25 \) when applied to the special case of an ideal Gaussian likelihood (per §2.2). Such a ladder will not, in general, give uniform acceptance ratios when applied to an arbitrary posterior distribution, but this choice reflects the more realistic scenario where we cannot guess at...
$T_{\text{prior}}$, and so we resort to assuming that the distribution indeed behaves like an ideal Gaussian.

Figure 13 shows the evolution of the temperatures and acceptance ratios for an emcee sampler of 15 chains under the temperature dynamics of [4]. Figure 14 shows the equilibrium density $\eta(\log T)$ after the ladder has achieved uniform acceptance ratios.

Figure 15 shows the ACTs of the cold chain ($T = 1$) under uniform-$A$ and geometric ladders for the 5-dimensional egg-box problem as a function of the number of temperatures available. In this case, adding more temperatures to a geometric ladder does not reduce the measured ACT of the sampler for $N \geq 7$, since they are added to the high-$T$ end of the ladder, above $T_{\text{prior}}$, and the ratios between lower temperatures do not change. Figure 15 shows that from the initial geometric ladder only around 6 chains are within the range of temperatures spanned by the equilibrium ladder; the remaining 8 (excluding $T_1 = 1$ and $T_N = \infty$) are all above $T_{\text{prior}}$ and effectively sample from the prior. In this case, therefore, the geometric spacing that would give uniform acceptance ratios of 25% for an ideal Gaussian in fact spaces temperatures too widely for $\geq 6$ chains.

Meanwhile, adding more chains to a dynamically adapted ladder clearly reduces the ACT of the sampler in this regime. Moreover, the ACT of a sampler using a uniform-$A$ ladder is always lower than that of a sampler using the geometric ladder of the same $N$. In the egg-box example, which requires a relatively close spacing of temperatures, the improvement is dramatic when many chains are used: $\tau_{\text{geo}} > 2\tau_{\text{acc}}$ for $N \geq 12$.

The failure of the geometric ladders used in this example for $N \geq 7$ lies in the poor $T_{\text{max}}$ chosen by assuming that the distribution behaves like an ideal Gaussian. A geometric spacing is in fact appropriate for a large portion of the temperature range, but its efficacy relies on the ladder terminating at the correct $T_{\text{prior}}$. When $N < 7$, the geometric and uniform-$A$ ladders show similar ACTs, and the geometric ladders in fact do slightly better. While unexpected, this is a consequence of the behaviour of the affine invariant ensemble sampler used in emcee (Foreman-Mackey et al. 2013; Goodman & Weare 2010) as applied to the egg-box likelihood [22]. When such a sampler is applied to a target distribution for which the number of modes $n_m$ is greater than the number of walkers $n_w$ used by the sampler, it behaves as though it is sampling from the prior (albeit inefficiently). There is therefore little benefit in having a chain sampling as high as $T_{\text{prior}}$, and so it is better – in terms of the ACT – to assign more chains to lower temperatures in order to increase their acceptance ratios. In our case, the egg-box likelihood has 243 modes in 5 dimensions, while the sampler uses only 100 walkers, so these walkers tend to become isolated from one another. Since the sampler relies on clustering of walkers on an individual mode to inform jump proposals within that mode, jumps are instead proposed between modes when there are on average fewer than one walker per mode.

We anticipate that running the same tests on a traditional single-walker MCMC sampler, or reducing the number of modes of the likelihood distribution so that $n_w \gg n_m$, will dramatically increase $\tau_{\text{geo}}/\tau_{\text{acc}}$ in the low temperature regime. We should expect that $\tau_{\text{geo}} \gg \tau_{\text{acc}}$ when $T_{\text{max}}(N) \ll T_{\text{prior}}$ for the geometric ladder and that $\tau_{\text{geo}} \approx \tau_{\text{acc}}$ when $T_{\text{max}}(N) \approx T_{\text{prior}}$.

Figure 15 therefore illustrates a very specific case for $N < 7$ that does not reflect the importance of choosing $T_{\text{max}} \approx T_{\text{prior}}$. Nonetheless, the ACTs of the two temperature allocation strategies – geometric and uniform-$A$ – are still fairly similar for $N \leq 7$ and there is a distinct improvement for $N > 7$.

5 DISCUSSION

The temperature selection scheme set out in [32] solves two problems in the application of parallel tempering:

(i) It identifies $T_{\text{max}} = \infty$ as a suitable temperature for the hot chain – such that it will sample from the prior – that is independent of the target distribution.

(ii) It allocates a fixed number of intermediate temperatures to give good communication between fixed extremal temperatures $T_{\text{min}}$ and $T_{\text{max}}$, and therefore efficient sampling of the target distribution – i.e. with few iterations between independent samples.

The intermediate temperatures are allocated so that acceptance ratios for swaps proposed between neighbouring pairs of chains are uniform across the temperature ladder. The dynamical algorithm that implements this scheme requires only two parameters: $\nu$ and $t_0$. These parameters, discussed in [32], describe only the initial dynamics of the temperatures, setting the time-scale for temperature adjustments, and do not determine the equilibrium uniform-$A$ ladder.

While a parallel tempering temperature configuration that is selected for uniform acceptance ratios between all chain pairs is not necessarily optimal in the ACT of the sampler, we have demonstrated that it is generally

Figure 14. Orange: The equilibrium density of chains per log $T$ for the egg-box run illustrated in Figure 13 where the acceptance ratios have settled to $A_1 \approx 0.65$. Blue: The square root of the specific heat for the egg-box distribution, calculated as described in Figure 6.

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
$\tau_{\text{geo}}$ (uniform-$A$) & $\tau_{\text{acc}}$ & $\sqrt{C_V}$ (arbitrary scale) \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\end{tabular}
\end{center}
better than a conventional geometric temperature configuration and provides more consistent behaviour across different likelihood distributions and numbers of chains. Importantly, the dynamics that achieve such a temperature ladder are simple and easily implemented, requiring very little tuning or intervention.

The factor by which the ACT is reduced by the uniform-A scheme depends strongly on the likelihood distribution that is explored and on the specific geometric ladder against which the uniform-A scheme is being compared. For a geometric ladder, one must make an ad hoc choice of the maximum temperature $T_{\text{max}}$; this is difficult and a poor guess can yield a very sub-optimal ladder. In particular, if $T_{\text{max}}$ is not high enough that the sampler can efficiently migrate between modes, then the ACT will be significantly higher than it needs to be. On the other hand, if $T_{\text{max}}$ is too high, then many of the chains are effectively sampling from the prior, and CPU time is being wasted in sampling from redundant tempered likelihood distributions.

The uniform-A temperature dynamics guarantee that, for a given number of chains $N$, no such wastage of CPU time occurs and that there will always be precisely one chain sampling at $T_{\text{max}} = \infty$ (i.e., sampling from the prior). Tests of the dynamics generally demonstrate lower ACTs when compared with geometric ladders of the same number of chains, $N$.

In [1.22] we demonstrated that, even with a judicious choice of $T_{\text{max}}$ that is close to $T_{\text{prior}}$, a traditional geometric ladder is outperformed by a ladder chosen for uniform acceptance ratios (with $T_{\text{max}} = \infty$). Figure [10] illustrates that, when $T_{\text{prior}}$ is known, a uniform-A ladder confers the greatest reduction in ACT when $N$ is small. In this case, the temperature ratio $\gamma$ of the geometric ladder is large enough that phase transitions in the distribution of log $L$ cause bottlenecks in the communication between hot and cold chains. The uniform acceptance scheme allocates more chains over these temperature regimes in an effort to optimise the communication.

For larger $N$, $\tau_{\text{geo}}/\tau_{\text{acc}} \approx 1$, suggesting that as long as there are no pairs of chains with prohibitively low swap acceptance ratios – a geometric spacing is adequate, as long as $T_{\text{max}}$ is chosen appropriately. It is not clear how the threshold for such a bottleneck in $A_i$ is determined, but it is likely related to the time-scale of the intra-chain motion of the sampler. If intra-chain jumps are accepted seldom with respect to the rate of inter-chain swaps, then increasing the inter-chain swap acceptance ratio is unlikely to make the sampler any more efficient.

In general, while $\tau_{\text{geo}} > \tau_{\text{acc}}$ for all $N$, the improvement fraction $\tau_{\text{geo}}/\tau_{\text{acc}}$ will asymptote to 1 as $N \to \infty$. The rate of decay will depend strongly on the target distribution. A system with a wide distribution of log $L$ (e.g., with many dimensions) or with sharp phase transitions at certain temperatures (e.g., with many modes of various shapes and weights) will see the most benefit from having many chains, while a better-behaved distribution without such features can be efficiently sampled with fewer.

Meanwhile, from our tests on the 5-dimensional egg-box distribution discussed in [1.13] we can see the consequences of a poor choice of $T_{\text{max}}$. While the egg-box distribution does not have as strong a phase transition as the double Rosenbrock function of [1.2] our ignorance of $T_{\text{prior}}$ means that a geometric ladder (which in this case is constructed from a fixed temperature ratio $\gamma$) is mostly worse than an uniform-A ladder. Figure [15] demonstrates

---

**Figure 13.** The evolution of temperatures $T_i$ and acceptance ratios $A_i$ while sampling with `emcee` from a 5-dimensional egg-box distribution, [22], with 15 chains. Chains 1 and 15 are not shown, having fixed temperatures $T_1 = 1$ and $T_{15} = \infty$. 
5.1 Evidence calculations

The current paper focuses mainly on the efficiency of a parallel tempered MCMC sampler in producing independent samples from its target distribution. Another important task in Bayesian statistical inference to compute the evidence integral of the posterior distribution. At a given temperature, this is given by

$$Z(\beta) \equiv \int L(\bar{\theta})^\beta p(\bar{\theta}) d\bar{\theta},$$

where $\beta \equiv 1/T$ is the inverse temperature.

Since we are interested in the untempered posterior, we wish to calculate $Z(1)$. Goggans & Chi (2004) use thermodynamic integration to calculate $Z$, showing that

$$\log Z(1) = \int_0^1 E[\log L]_\beta d\beta.$$  

The log evidence can therefore be computed by a sampler through numerical integration of the mean log $L$ values collected over all of the chains. In the same way that inter-chain communication is hindered by phase transitions in the system, numerical estimation of this integral is susceptible to sharp changes in log $L$ with the temperature $T$. Such phase transitions are marked by a diverging specific heat $C_V$ since, from $\frac{1}{2}T^2$, $C_V$ is the derivative of log $L$ with respect to $T$.

Since allocating temperatures for uniform acceptance ratios yields a logarithmic chain density $\eta$ that appears to scale with $\sqrt{C_V}$, such a temperature ladder will naturally increase the accuracy of numerical estimates of $\frac{1}{2}T^2$ with respect to one that does not increase $\eta$ around phase transitions.

5.2 Other measures of optimality

We have investigated the performance of a temperature ladder adapted for uniform acceptance ratios in reducing the ACT of a parallel tempered MCMC sampler. The KL divergence discussed in 4.2 provides an alternative measure of the distance between two temperatures. In 4.1, we showed that uniform KL divergence in a temperature ladder does not correspond to uniform acceptance ratios beyond the special case of the ideal, unbounded Gaussian distribution described in 4.2.

When applied to the truncated Gaussian discussed in 4.1 for which $D_{\text{KL}}(\pi_{T_i} || \pi_{T_{i+1}})$ is analytically available, the $D_{\text{KL}}$ and $A$ measured between chains drop off at different rates as $T$ approaches $T_{\text{prior}}$ (see Figure 5). Nonetheless, there is no obvious reason to favour uniform $D_{\text{KL}}$ over uniform $A$ as a criterion for optimality. Since there is no practical way of computing the KL divergence while sampling, we settle for uniform acceptance ratios as our optimality criterion.

Meanwhile, Katzgraber et al. (2006) propose an optimisation scheme in which temperatures are chosen to minimise the round-trip time of a sample from $T_{\text{min}}$ to $T_{\text{max}}$, which they suggest will improve sampling performance on systems with strong phase transitions. Their algorithm is tested on simulations of the two-dimensional

Figure 15. Top: the ACTs of the cold chain ($T = 1$) of a sampler exploring the egg-box likelihood (22) with ladders of different sizes $N$, for both geometric temperature ladders and ladders dynamically adapted for uniform acceptance ratios. Middle: the total CPU time, $N \times \tau_i$, for the runs. Bottom: the relative improvement in ACT conferred by dynamically adapting for uniform acceptance ratios over a geometric ladder.

dthis, specifically when $N$ is large enough that for a given $\gamma$ the geometric ladder places many temperatures redundantly above $T_{\text{prior}}$. In this case, we see a dramatic improvement in ACT $\tau$ from using a uniform-$A$ ladder when compared with a geometric ladder of the same number of chains $N$; indeed, the ratio $\tau_{\text{geo}} / \tau_{\text{acc}}$ becomes as large as $\sim 4$ for the values of $N$ tested. Since $\tau_{\text{geo}}$ is independent of $N$ when $N \geq 7$, we should expect that this ratio will saturate as $N \to \infty$, where $\tau_{\text{acc}}$ reaches a minimum. Moreover, the CPU time, $N \times \tau$ of the uniform-$A$ runs continues to decrease with $N$, even as the CPU time of the geometric runs rises.

On the other hand, when $N$ is too small for a geometric ladder to reach the prior (i.e., $T_N \ll T_{\text{prior}}$), we notice that in fact $\tau_{\text{geo}} < \tau_{\text{acc}}$. As discussed in 4.3, this somewhat surprising result arises from a limitation of the ensemble sampler that was used to sample the distribution. We anticipate that, if either a traditional single-walker sampler were used or the number of walkers were increased to many times the number of modes, the geometric ladder will fail dramatically in this regime of $N$, giving $\tau_{\text{geo}} \gg \tau_{\text{acc}}$. 

\begin{equation}
\frac{\tau_{\text{geo}}}{\tau_{\text{acc}}}
\end{equation}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure15}
\caption{Top: the ACTs of the cold chain ($T = 1$) of a sampler exploring the egg-box likelihood (22) with ladders of different sizes $N$, for both geometric temperature ladders and ladders dynamically adapted for uniform acceptance ratios. Middle: the total CPU time, $N \times \tau_i$, for the runs. Bottom: the relative improvement in ACT conferred by dynamically adapting for uniform acceptance ratios over a geometric ladder.}
\end{figure}
Temperature dynamics for PTMCMC samplers

Ising model, and is shown to select a different temperature configuration than the uniform-A scheme that has been discussed so far.

However, the ACTs of the sampler — what we are ultimately concerned with in Bayesian inference — is not discussed, so it is unclear whether this strategy is better than selecting temperatures for uniform acceptance ratios. Their feedback optimisation method in fact prefers a higher density of chains per $T$ across phase transitions of the system than the uniform-A scheme. We have shown, however, that the ACT yielded by a particular ladder is not critically sensitive to under-densities over phase transitions so long as the acceptance ratio is not prohibitively small in these temperature regimes (see §4.2.2). Indeed, increasing the density of chains over phase transitions too far might unnecessarily hinder inter-chain communication at other temperatures (by reducing $A$), leading to an overall drop in ACT.

These reservations, together with the complicated book-keeping involved in optimising for round-trip time, lead us to favour the dynamical method presented in §3. By comparison, this dynamical method is simple and guaranteed to produce an equilibrium ladder that yields efficient — if not perfectly optimal — sampling from any target distribution.

6 SUMMARY

Selecting temperatures is an important open problem in all applications of parallel tempering. Temperatures must be selected so that each phase of the tempered target distribution is accessible to at least one chain of the sampler, while maintaining sufficient communication between chains that the chain at the temperature of interest benefits from the parallel exploration of chains at other temperatures.

The key results of this paper are two-fold. Firstly, we have proposed that temperatures be selected such that swap proposals between neighbouring chains are accepted with a probability that is uniform across the temperature ladder. Secondly, in §4, we have developed a dynamical algorithm for achieving such a temperature distribution on-the-fly, as the sampler is running.

Tests in §4 show that, while a uniform-A temperature ladder is not necessarily optimal, the prescription consistently gives efficient sampling — i.e. few iterations between independent samples — regardless of the distribution that is being explored.

Importantly, the dynamical algorithm of §4 that generates such a temperature ladder needs very little tuning to work effectively. It requires only two parameters (see §4.2), which determine the transient dynamics that lead to the uniform-A ladder. Moreover, it is conceptually simple and implementation requires little modification of existing code. A reference implementation of the algorithm, on the ensemble sampler emcee (Foreman-Mackey et al. 2013), is available at https://github.com/williamshen/emcee/tree/adaptive

Most parallel tempered MCMC sampling packages require an ad-hoc temperature specification that presumes prior knowledge of the distribution to be sampled. The dynamical algorithm presented in this paper automates the selection of temperatures in a way that works out-of-the-box with little to no tuning required, reliably achieving good sampling performance independently of the target distribution.

The simplicity of this method, and its ease of use and implementation, make it an asset to astronomers wishing to investigate difficult, high-dimensional Bayesian inference problems.

ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Facilities Council and a Leverhulme Trust research project grant.

REFERENCES

Earl D. J., Deem M. W., 2005, Phys. Chem. Chem. Phys., 7, 3910
Falcioni M., Deem M. W., 1999, J. Chem. Phys., 110, 1754
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, Pub. Astron. Soc. Pac., 125, 306
Geyer C. J., 1991, in Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface Markov chain Monte Carlo maximum likelihood. Interface Foundation of North America, New York, pp 156–163
Goggans P. M., Chi Y., 2004, AIP Conf. Proc., 707, 59
Goodman J., Weare J., 2010, Commun. Appl. Math. Comp. Sci., 5, 65
Hukushima K., Nemoto K., 1996, J. Phys. Soc. Jpn., 65, 1604
Katzgraber H. G., Trebst S., Huse D. A., Troyer M., 2006, J. Stat. Mech: Theory Exp., 2006, P03018
Koike D. A., 2002, J. Chem. Phys., 117, 6911
Koike D. A., 2004, J. Chem. Phys., 120, 10852
Rathore N., Chopra M., de Pablo J. J., 2005, J. Chem. Phys., 122, 1604
Roberts G. O., Rosenthal J. S., 1998, Can. J. Stat., 26, 5
Roberts G. O., Rosenthal J. S., 2007, J. App. Prob., 44, 458
Sanbonmatsu K., Garcia A., 2002, Proteins: Struct., Funct., Bioinf., 46, 225
Schug A., Herges T., Wenzel W., 2004, Proteins: Struct., Funct., Bioinf., 57, 792
Sokal A., 1997, in DeWitt-Morette C., Cartier P., Folacci A., eds, NATO ASI Series, Vol. 361, Functional Integration. Springer US, pp 131–192
Sugita Y., Okamoto Y., 1999, Chem. Phys. Lett., 314, 141
Swendsen R. H., Wang J.-S., 1986, Phys. Rev. Lett., 57, 2607
ter Braak C., Vrugt J., 2008, Stat. Comput., 18, 435

© 2002 RAS, MNRAS 000
APPENDIX A: AUTOCORRELATION TIME ESTIMATION

The autocorrelation time (ACT) discussed in this paper refers to the integrated autocorrelation time described by Sokal (1997). It is estimated in the following way.

If \( x(t) \) is a time series with a normalised autocorrelation function \( \rho(t) \), such that \( \rho(0) = 1 \), then the integrated ACT of \( x \) is defined by

\[
\tau \equiv \sum_{t=-\infty}^{\infty} \rho(t)
\]

\[
= 1 + 2 \sum_{t=1}^{\infty} \rho(t).
\]

Since, when \( t \gg \tau \), \( \rho(t) \approx 0 \), there is little contribution to the integral at large lags, except through noise in the measured autocorrelation function \( \rho \). We can therefore express this relation through the consistency condition

\[
\tau \approx 1 + 2 \sum_{t=1}^{\infty} \rho(t).
\]

In words, we estimate the ACT over a window that is \( M \) ACTs long, subject to the constraint that \( M \tau < N/2 \), if \( N \) is the number of samples in \( x \). If this constraint is violated, the result is probably not trustworthy, since there are too few samples for a meaningful estimate.