Tuning Tempered Transitions

Gundula Behrens[*], Nial Friel[†] and Merrilee Hurn[‡]

October 6, 2010

Abstract

The method of tempered transitions was proposed by Neal (1996) for tackling the difficulties arising when using McMC to sample from multimodal distributions. In common with methods such as simulated tempering and Metropolis -coupled MCMC , the key idea is to utilise a series of successively easier to sample distributions to improve movement around the state space. Tempered transitions does this by incorporating moves through these less modal distributions into the MCMC proposals. Unfortunately the improved movement between modes comes at a high computational cost with a low acceptance rate of expensive proposals. We consider how the algorithm may be tuned to increase the acceptance rates for a given number of temperatures. We find that the commonly assumed geometric spacing of temperatures is reasonable in many but not all applications.

Keywords: Markov Chain Monte Carlo, Multimodality, Tempering, Thermodynamic Integration.

1 Introduction to tempering ideas in MCMC

It is well known that standard MCMC methods , such as the Metropolis - Hastings algorithm or the Gibbs sampler, often have difficulties in moving around their target distribution. When a chain mixes poorly in this way, there is a danger that modes have been missed or that modes are not represented in their right proportions, both of which may lead to bias in the statistical inference. To overcome such mixing problems, various more sophisticated MCMC methods have been devised based on a few key ideas. This paper concentrates on one of these key ideas, namely tempering.

One way to motivate TIS is to think of using to estimate some expectation $\mathsf{E}_{p_0}[h(X)]$ wrt the target distribution p_0 by sampling from some less modal distribution p_1.

[*]Department of Epidemiology and Preventive Medicine, University Hospital Regensburg 93042 Regensburg, Germany; Gundula.Behrens@klinik.uni-regensburg.de

[†]School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Republic of Ireland; Email: nial.friel@ucd.ie

[‡]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; Email: M.A.Hurn@bath.ac.uk

Asymptotically flatten distributions

One possibility for generating a less modal distribution than p_0 on the same support is to "flatten" it by taking $p_1(x) \propto p_0(x)^\beta$, $\forall x$, with $\beta < 1$. As $\beta \to 0$, $p_1(x)$ becomes closer to a uniform distribution and consequently becomes more amenable to sampling. For β close to 1, there is far less benefit as p_1 may not be that much less modal than p_0. Unfortunately as the two distributions become far enough apart that the difficulty with modality is overcome, they may also become far enough apart so that many of the importance weights will be very close to zero resulting in unstable estimates of $\mathsf{E}_{p_0}[h(X)]$. The basis of all the tempering methods is the introduction of a series of distributions bridging the gap between p_0 and p_1. The differences between the various approaches is in how these bridging distributions are included. We will describe various approaches to incorporating bridging distributions using the common form of tempering which involves powering up all or part of the unnormalised target distribution. The inclusion of other types of bridging distribution would also be possible, but the literature has generally restricted itself to this form

the target distribution :

$$p(x) \propto \pi(x)\, \exp(-\beta_0\, h(x)), \tag{1}$$

where $h(x)$ may be known as the "energy" function and the parameter β_0 as the target "inverse temperature". The tempered distributions:

$$p_i(x) \propto \pi(x)\, \exp(-\beta_i\, h(x)), \qquad i = 0, 1, \ldots, n, \tag{2}$$

where $0 \leq \beta_n \leq \ldots \leq \beta_1 \leq \beta_0$, are the inverse temperatures characterising each distribution. The flexibility of potentially only tempering part of the target distribution is quite useful. In Bayesian problems it may be that one or other of the prior and likelihood contribute to the mixing problems.

One of the earlier suggestions for incorporating tempering into MCMC is to run $n+1$ Markov chains in parallel, each sampling from one of the $n+1$ tempered distributions. At each iteration, proposals are made to update each chain separately and additionally there is a proposal to swap the x values between chains thereby coupling them and giving rise to the name Metropolis-coupled MCMC (Geyer 1991). The state space is the enlarged set of $(n+1)$ values for x and the target distribution is $p_0 \otimes p_1 \otimes \ldots \otimes p_n$. The idea is that large moves made under p_n will filter back down to the lowest level p_0. The normalising constants for the tempered distributions are not needed in this method as they appear only in the acceptance probabilities for the coupling move where they cancel out. However the tempered distributions do need to be close in order that the swaps between them are not too infrequent. This may mean that n will have to be large and there are then obvious consequences for storage and computational effort.

A single chain alternative to Metropolis coupling is simulated tempering (Marinari and Parisi 1992,

Geyer and Thompson 1995) which runs a chain on the state space of x augmented by a variable i which takes the values $i = 0, 1, \ldots, n$ with probabilities determined by a "pseudo-prior". The stationary distribution of $x|i$ is $p_i(x)$, and updates are either of $x|i$ or $i|x$ with the latter effectively moving up or down the tempering sequence. Although again the normalising constants of the tempering distributions are not needed explicitly, in practice to get reasonable acceptance rates for the moves between temperatures, the pseudo-prior needs to be roughly proportional to these unknown normalising constants.

Tempered transitions (Neal 1996) is another single chain method but without the need to guesstimate the relative normalising constants of the tempered distributions. It uses a deterministically ordered sweep up and down the tempering distributions as a way of generating proposals for the main chain in a way which will be described in more detail in the next section. The overwhelming cost of the algorithm is in the construction of the proposals and therefore it is imperative that these are tuned carefully to maximise acceptance rates. Neal (1996) finds tempered transitions and simulated tempering to be of roughly equal computational cost. He also compares tempered transitions, simulated tempering and Metropolis-coupled MCMC on other criteria such as storage requirements and the number of tempering levels required concluding that there is no overall winner and that the choice of method may be problem and goal specific.

Closely related methods which aim to make fuller use of the sampling at all temperature levels via importance sampling can be found in Neal (2001) and more recently in Gramacy, Samworth and King (2010). The former has links to tempered transitions, effectively using just the second half of the complex proposal mechanism. The latter has stronger connections with simulated tempering and Metropolis coupled MCMC where samples from the different temperatures are stored. Other instances of ideas involving populations of samples can be found in both the population-based MCMC and the Sequential Monte Carlo literature (see Jasra, Stephens and Holmes (2007) for an overview).

A common question arising across the algorithms involving tempering is the choice of the bridging distributions given by Equation (2). The general recommendation is to space the βs geometrically, that is so that β_i/β_{i+1} is a constant for all levels i (Neal 1996). Neal formulates this rule by considering sampling from a multivariate Gaussian using simulated tempering; the geometric spacing attempts to maximise the acceptance rates of swaps between neighbouring chains at all levels. Gelman and Meng (1998) also consider choices of bridging distribution, although in the context of the closely related question of estimating normalising constants where they are trying to minimise the Monte Carlo error of path sampling estimates. Other work on rationales for choosing the βs can be found in Iba (2001) and Lefebvre, Steele and Vandal (2010). The former reviews the (largely physics) literature, comparing simulated tempering with exchange and ensemble Monte Carlo methods and aims to maximise the swapping rates between the bridging distributions using preliminary runs (to satisfy a theoretically derived optimality criterion). The latter is interested

in path sampling for estimating normalising constants and the related tuning of the bridging distributions; they derive an expression for the symmetrised K-L divergence between pairs of distributions and use the minimisation of this as their criterion.

In this paper we consider tempered transitions with bridging distributions of the form given by Equation (2). Of the various tempering schemes, the choice of the $\{\beta_i\}$ seems to have been least well addressed for tempered transitions. Our approach is largely computational and tries to answer the question "For a given number n of tempering distributions, how best should they be spaced?". We note that the question of how we should choose n, for fixed computational time, is beyond the scope of this article. The approach we take is to use a small number of preliminary short runs to assess whether geometric spacing is likely to be adequate and, if not, we propose an optimal way of spacing them. In Section 2 we describe tempered transitions in detail, building on many of the insights offered in Neal's paper. We provide a theoretical analysis which outlines when geometric temperature spacing is optimal and give a motivating example where geometric spacing is sub-optimal. We also draw some parallels with some of the other theoretical approaches outlined above. In Section 3 we discuss the implementation details of applying our proposed approach to a slowly mixing MCMC application.

2 How to tune tempered transitions?

2.1 The tempered transitions algorithm

We begin by describing the algorithmic details of the tempered transitions algorithm and setting up the reasoning behind our tuning approach. Suppose the chain is currently in state x, then the algorithm generates a proposal x' for the next state using a secondary chain which passes through all the auxiliary distributions $\{p_i\}$ first in ascending order of the βs ("heating-up") and then in descending order ("cooling-down") back to the target distribution p_0. To do this, it uses n pairs of MCMC transition kernels with the i^{th} pair, T_i and T_i' satisfying detailed balance wrt p_i

$$p_i(x)\,T_i(x,x') = p_i(x')\,T_i'(x',x) \quad \forall x, x', i = 1, \ldots, n.$$

Step 1 Set $x_0 = x$.

Step 2 Move up and down the tempered distributions using MCMC transitions

Generate x_1 from x_0 using transition kernel T_1.

Generate x_2 from x_1 using transition kernel T_2.

$\vdots$

Generate x_n from x_{n-1} using transition kernel T_n.

Generate x'_{n-1} from x_n using transition kernel T'_n.

$\vdots$

Generate x'_1 from x'_2 using transition kernel T'_2.

Generate x'_0 from x'_1 using transition kernel T'_1.

Step 3 Accept $x' = x'_0$ as the next state with probability

$$\alpha(x, x'|x_0, x_1, \ldots, x_n, x'_{n-1}, \ldots, x'_1, x'_0) = \min\left\{1, \left[\prod_{i=0}^{n-1} \frac{p_{i+1}(x_i)}{p_i(x_i)}\right] \left[\prod_{i=0}^{n-1} \frac{p_i(x'_i)}{p_{i+1}(x'_i)}\right]\right\}, \qquad (3)$$

otherwise, remain at state x.

There is no need for the normalising constants of the tempering distributions as they cancel in the acceptance probability. Neal (1996) demonstrates that the algorithm satisfies detailed balance wrt the target p_0. However it is perhaps clear that the proposal is potentially computationally costly and tricky to tune.

There are two dependent aspects to the tuning: 1 the $\{T_i, T'_i\}$ should have reasonable acceptance rates and, at the later stages of the tempering, be able to make large moves in the state space (otherwise this expensive proposal makes little change). This is not quite the usual tuning problem for MCMC in that each successive level has a different target distribution. During the first half, these distributions are becoming progressively less modal, while in the second half the reverse is true. Obviously the closer the consecutive distributions, ie the closer the consecutive βs, the less of an effect this will be.

2 Subject to the individual $\{T_i, T'_i\}$ working well, the overall acceptance rate of the entire tempered transition proposal should be as high as possible (although notice that if all the proposed changes in the tempering are rejected, the overall proposal will be accepted since $x_0 = x_1 = \ldots = x'_1 = x'_0$ and so a high acceptance rate here can be slightly misleading if viewed in isolation). To gain more insight into tuning the tempered transition acceptance rate, we follow Neal's lead and rewrite the acceptance probability, Equation (3), using the form of tempering defined by Equation (2).

$$\begin{aligned}
\alpha(x, x'|x_0, x_1, \ldots, x'_1, x'_0) &= \min\left\{1, \left[\prod_{i=0}^{n-1} \frac{\exp(-\beta_{i+1}h(x_i))}{\exp(-\beta_i h(x_i))}\right] \left[\prod_{i=0}^{n-1} \frac{\exp(-\beta_i h(x'_i))}{\exp(-\beta_{i+1}h(x'_i))}\right]\right\} \\
&= \min\left\{1, \exp(-(F' - F))\right\} \qquad (4) \\
\text{where } F &= \sum_{i=0}^{n-1} (\beta_i - \beta_{i+1}) h(x_i) \\
\text{and } F' &= \sum_{i=0}^{n-1} (\beta_i - \beta_{i+1}) h(x'_i).
\end{aligned}$$

This expression has an interpretation related to estimating the ratio of normalising constants by thermodynamic integration. Let $Z(\beta)$ denote the normalising constant of the distribution defined by Equation (2),

where for the moment we assume that β takes continuous values in the interval $[\beta_0, \beta_n]$.

$$\begin{aligned} Z(\beta) &= \int_x \pi(x) \exp(-\beta h(x))\, dx \\ \text{then} \quad \frac{dZ(\beta)}{d\beta} &= -Z(\beta)\mathsf{E}_\beta[h(X)]. \end{aligned} \tag{5}$$

Solving this differential equation for $Z(\beta)$ gives

$$\log(Z(\beta_n)) - \log(Z(\beta_0)) = \int_{\beta_n}^{\beta_0} \mathsf{E}_\beta[h(X)]\, d\beta. \tag{6}$$

That is, the log of the ratio of the normalising constants at two values of β can be expressed as the area under the curve $g(\beta) = \mathsf{E}_\beta[h(X)]$ between them. Recall that the sequences $\{x_0, \ldots x_{n-1}\}$ and $\{x'_{n-1}, \ldots x'_0\}$ are drawn such that the x_i have target distribution p_i, while the x'_i have target distribution p_{i+1}. Figure 1 illustrates a slightly idealised realisation of F (left) and F' (right) as the shaded areas constructed as the sum of rectangles with widths $(\beta_i - \beta_{i+1})$ and heights $h(x_i)$ (left) or $h(x'_i)$ (right). Both areas are approximations of the integral of $g(\beta)$ between β_n and β_0. Different realisations of $\{x_0, x_1, \ldots, x'_0\}$ will obviously give quite different and usually rather messier pictures, with correspondingly quite varied values of $F' - F$. (In fact, Figure 1 was constructed using the average of several realisations to reduce this variability for presentation purposes.) Those realisations of $x_0, x_1, \ldots, x'_1, x'_0$ where the shaded area on the right (F') is smaller than that on the left (F) will be accepted since in that case $\exp(-(F' - F)) > 1$ in Equation (4). Those for which F' is slightly larger than F may be accepted, but we will almost certainly reject those for which $(F' - F)$ is large. We take this as a motivation for selecting the $\{\beta_i\}$ for fixed n.

2.2 The proposed rationale for choosing $\{\beta_i\}$

Given the cost of each tempered transition proposal, our motivation is to increase the number of proposals accepted. The value of β_0 is fixed by Equation (1), and we assume that the other extreme of the βs is also determined, possibly by the fact that it defines a distribution for which direct sampling is possible, certainly by the need to move around the state space freely under p_n. What remains undetermined are n and the set $\{\beta_1, \ldots, \beta_{n-1}\}$.

Figure 1 showed the F and F' associated with a realisation $\{x_0, x_1, \ldots, x'_0\}$. If at each stage the transitions were able to reach their equilibrium distributions in the one step available, then $\mathsf{E}[h(x_i)] = g(\beta_i)$ and $\mathsf{E}[h(x'_i)] = g(\beta_{i+1})$; Figure 2 shows the corresponding approximations to the integral of $g(\beta)$ (this is equivalent to Neal's Figure 1(a)). Denote this difference between the areas of these two step functions as a

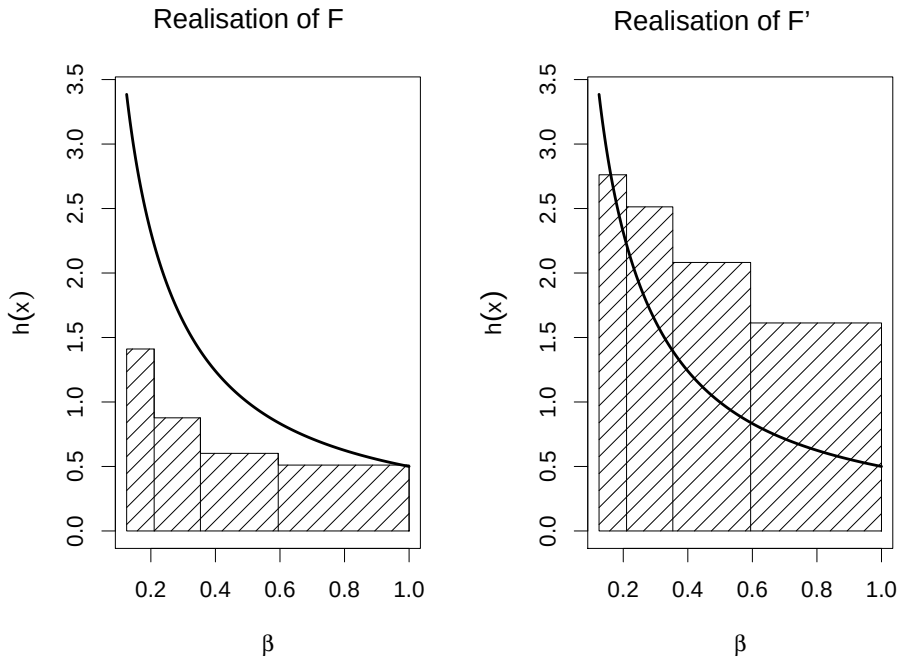

Figure 1: Two approximations to the integral of $g(\beta)$. The breakpoints of the rectangles are given by $\beta_n \leq \ldots \leq \beta_1 \leq \beta_0$. The shaded area on the left is $F = \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})h(x_i)$, while that on the right is $F' = \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})h(x_i')$. The overlaid curve is $g(\beta) = \mathsf{E}_\beta[h(X)]$.

function of the β values

$$\begin{aligned} S_n(\beta_0, \ldots, \beta_n) &= \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})\mathsf{E}_{i+1}[h(X)] - \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})\mathsf{E}_i[h(X)] \\ &= \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})g(\beta_{i+1}) - \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})g(\beta_i). \end{aligned} \tag{7}$$

Some results are well known for $g(\beta) = \mathsf{E}_\beta[h(X)]$. Rewriting Equation (2) as

$$p_\beta(x) = \pi(x)\ \exp(-\beta\, h(x) - K(\beta)) \tag{8}$$

so that $K(\beta)$ is the log of the normalising constant $Z(\beta)$ and rearranging Equation (5),

$$\begin{aligned} g(\beta) &= -\frac{1}{Z(\beta)}\frac{dZ(\beta)}{d\beta} \\ &= -\frac{d\log(Z(\beta))}{d\beta} \\ &= -K'(\beta) \end{aligned} \tag{9}$$

$$\text{and } g'(\beta) = \int h(x)\frac{d}{d\beta}p_\beta(x)dx$$

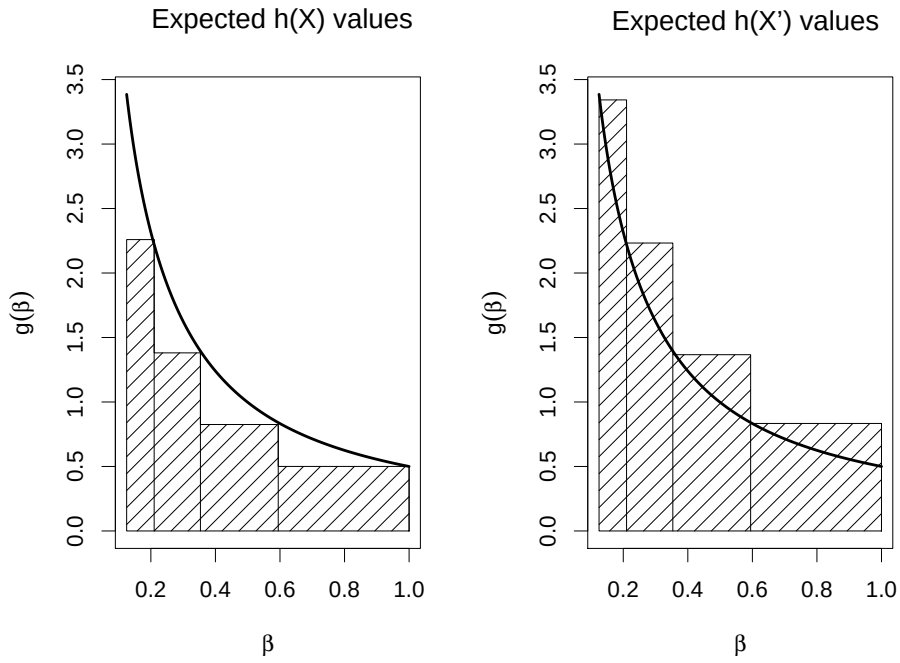

Figure 2: The shaded area on the left is $\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})g(\beta_i)$, while that on the right is $\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})g(\beta_{i+1})$. The difference between the two areas is $S_n(\beta_0, \ldots, \beta_n)$.

$$\begin{aligned} &= \int h(x)(-h(x) - K'(\beta))\pi(x)\exp(-\beta\, h(x) - K(\beta))dx \\ &= \int (-h(x)^2 + h(x)g(\beta))p_\beta(x)dx \\ &= -\mathrm{Var}_\beta\left[h(X)\right]. \end{aligned} \tag{10}$$

Therefore $g'(\beta) < 0$, for all β, showing that $g(\beta)$ is a decreasing function of β. It is possible to examine $g''(\beta)$ similarly, showing that the curve may be convex, concave, or a mixture of the two. The main point here is that because $g(\beta)$ is decreasing, we know that $S_n(\beta_0, \ldots, \beta_n) \geq 0$.

We propose the minimisation of $S_n(\beta_0, \ldots, \beta_n)$ over $\{\beta_i\}$ as our rule for choosing the tempered transition parameters. Obviously increasing n immediately reduces S_n. However our primary motivation here is the most effective choice of the particular $\{\beta_1, \ldots, \beta_{n-1}$ for a fixed number of levels n and fixed values of β_0 and β_n.

Note that minimising $S_n = \mathsf{E}[F' - F]$ is not directly equivalent to maximising the expected value of the acceptance probability, $\alpha = \min\{1, \exp(-(F' - F))\}$, however

$$\mathsf{E}[\exp(-(F' - F))] = 1 - S_n + \frac{\mathsf{E}[(F' - F)^2]}{2!} - \ldots \tag{11}$$

and so intuitively minimising S_n seems a reasonable start. Other possible criteria include, for example,

maximising $\mathsf{P}(F' < F)$ over the $\{\beta_i\}$ or, as suggested by one of the referees, examining the variance as well as the expectation of $F' - F$ since a high variance could perhaps improve mixing by generating big moves more often than a low variance might. We have so far only considered looking at S_n.

2.3 A motivating example

To motivate the tuning procedure, we study the one-dimensional two-parameter simplified Witch's Hat distribution used by Geyer and Thompson (1995). Although this is quite a straightforward example, we shall see that it is one for which geometric spacing of the temperatures is not optimal. Geyer and Thompson attribute this distribution to Matthews (1993) who introduced it as a problem case for the Gibbs sampler:

$$p(x) \propto 1 + b\mathsf{I}_{[x \leq a]}, \quad 0 \leq x \leq 1$$

where the parameters satisfy $0 < a < 1$ and $b \geq 0$. This apparently innocuous L-shaped distribution causes problems for standard Metropolis-Hastings moves if a is small but b is large as it can be difficult to move between the intervals $[0, a]$ and $(a, 1]$. The distribution can be expressed in a form suitable for tempering

$$p_i(x) \propto \exp(\beta_i \log(1 + b\mathsf{I}_{[x \leq a]})), \quad i = 0, \ldots, n$$

where $\beta_0 = 1$. In this case, $g(\beta) = \mathsf{E}_\beta[-\log(1 + b\mathsf{I}_{[x \leq a]})]$ is available analytically as are its derivatives:

$$\begin{aligned} g(\beta) &= \frac{-a(1+b)^\beta \log(1+b)}{a(1+b)^\beta + (1-a)} \\ g'(\beta) &= \frac{a(a-1)(1+b)^\beta (\log(1+b))^2}{(a(1+b)^\beta + 1 - a)^2} \\ g''(\beta) &= \frac{-a(a-1)(1+b)^\beta (\log(1+b))^3}{(a(1+b)^\beta + (1-a))^3} \left(a(1+b)^\beta - (1-a)\right) \end{aligned} \tag{12}$$

The second derivative shows that for $\beta \in [0, 1]$, the curve $g(\beta)$ may be convex (if $a \in [0.5, 1]$ and $b \geq 0$), concave (if $a \in (0, 0.5)$ and $b \in (0, 1/a - 2]$), or a mix of the two (otherwise). We propose to study one distribution for which $g(\beta)$ is convex (taking $a = 0.5$ and $b = 7.5 \times 10^8$) and one for which it is concave (taking $a = 10^{-4}$ and $b = 9.5 \times 10^3$). The latter distribution is hard to sample with roughly half the mass concentrated in the narrow $[0, 9.5 \times 10^3]$ peak. The former distribution does not present a sampling problem, but we are still interested in the effect of the shape of $g(\beta)$ on tempering performance.

Given $g(\beta)$, β_0, β_n and n, how do we minimise S_n over $\{\beta_1, \ldots, \beta_{n-1}\}$, Equation (7)? The $(n-1)$ partial derivatives are available,

$$\frac{\partial S_n}{\partial \beta_i} = (g(\beta_{i-1}) - 2g(\beta_i) + g(\beta_{i+1})) + (\beta_{i-1} - 2\beta_i + \beta_{i+1}) g'(\beta_i), \quad i = 1, \ldots, n-1, \tag{13}$$

however, despite their relatively simple form, no analytic solution is readily available for the minimisation problem. As a result, we perform the minimisation numerically using the built-in quasi-Newton optim with

option L-BFGS-B in R incorporating derivative information and the constraints that $\beta_n < \beta_i < \beta_0$ for $i = 1, \ldots, n-1$ and fixed β_0 and β_n. This function is not guaranteed to converge to a global maximum; in our experiments it was insensitive to starting points (we used geometric or uniform spacings) with the exception of occasional catastrophic convergence for large n to a non-ordered set of βs. In all cases we encountered, changing from geometric to uniform initial spacings, or vice versa, resolved this problem. Figure 3 illustrates the minimal S_n and the corresponding $\{\beta_i\}$ for the two examples of the Witch's hat distribution when $n = 4$ and $\beta_n = 1/16$, as well as the equivalent S_n for a geometric spacing of the $\{\beta_i\}$. Unsurprisingly given the different shapes of the two curves, the change in the size of S_n achieved by the optimal scheme over the geometric one is more significant for the concave $g(\beta)$ curve. However, even in the convex example it is clear that the values of the βs themselves, particularly β_1, are quite different under geometric spacing and the minimal S_n scheme. Table 1 shows the minimum values of S_n and the geometric values of S_n for $n = 2$, 4, 8, 16, 32 and 64 and for both pairs of parameters a and b. For each n we performed 500000 iterations of tempered transitions where at each level $i = 1, \ldots, n$ we use direct sampling (by inversion) to draw from the tempered distribution p_i. This direct sampling is only realistically possible, at least for values of β close to β_0, in a test example such as this, but it does allow us to separate the effects of the different β choices from the effects of slow mixing of the transitions at levels 1 to n. Table 1 gives observed average acceptance rates together with the estimated integrated autocorrelation time of the tempering calculated with respect to the known theoretical mean. The Witch's Hat example is unusual in that only moves between the regions $0 \leq x \leq a$ and $a < x \leq 1$ are problematic while all within-region moves are always accepted (giving rise to unusually high acceptance rates for a tempering problem). In addition, a high acceptance rate in tempered transitions can actually mask a lack of mixing and so integrated autocorrelation times are a useful diagnostic.

For both distributions, decreasing S_n by optimising the $\{\beta_i\}$, increases the observed acceptance rate and decreases the integrated autocorrelation time. These improvements are most noticeable when comparing the geometric and optimal schemes for the hard sampling problem, $a = 10^{-4}$ and $b = 9.5 \times 10^3$, where the changes in S_n are most dramatic. Concentrating on $n = 4$ to tie in with Figure 3, for the easier convex $g(\beta)$, optimising the $\{\beta_i\}$ made a small difference to the overall S_n albeit with noticeable changes to the $\{\beta_i\}$ themselves. Here the tuning has made only marginal improvements in acceptance rates and integrated autocorrelation times (although as noted earlier, this distribution is not hard to sample and there is little scope for improvement anyway). In the harder sampling problem, where $g(\beta)$ was concave, the benefits of tuning the $\{\beta_i\}$ are very clear. In this example, the additional computational cost of tuning comes only from the R optimisation stage. The benefits of tuning are greatest when n is small (as n increases, the geometric S_n anyway decreases to zero).

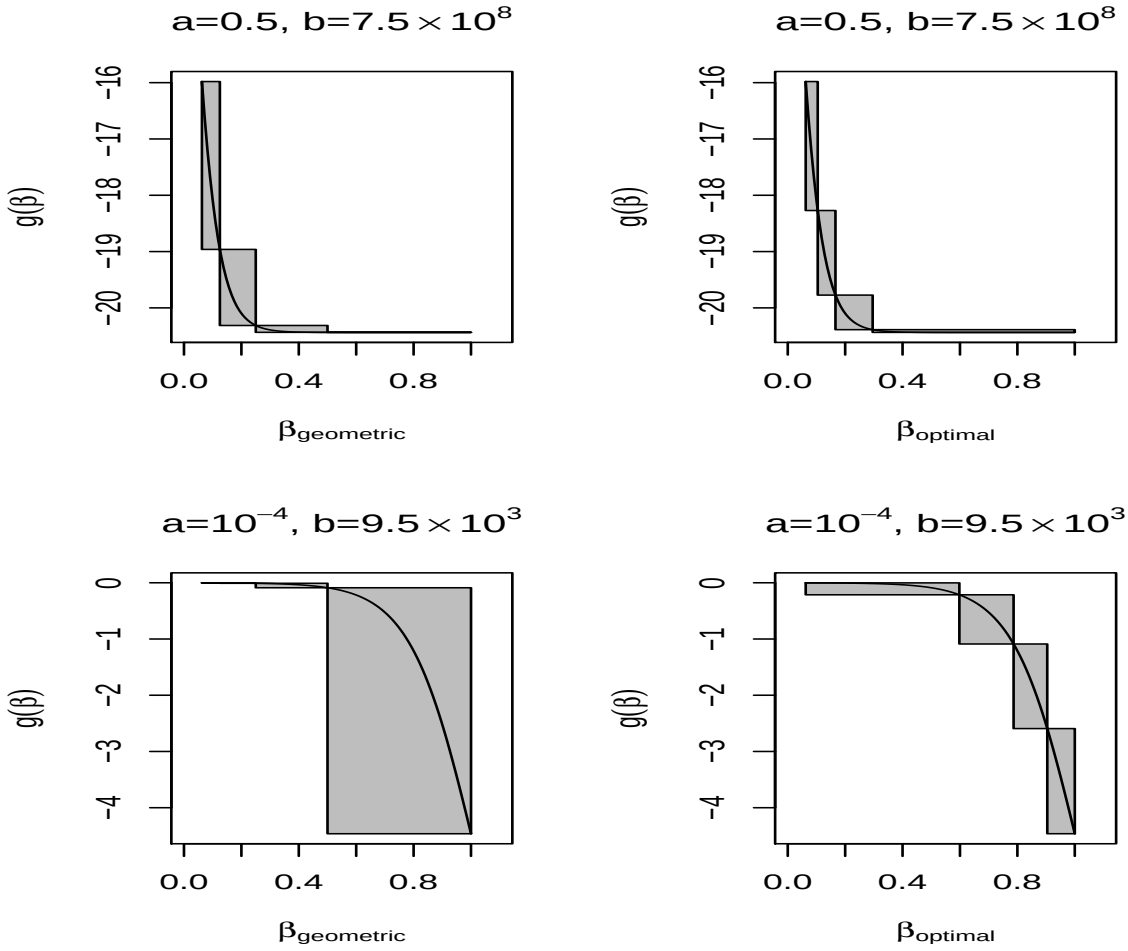

Figure 3: S_n for the Witch's hat distribution when $n = 4$ using geometric $\{\beta_i\}$ spacing (left) or the optimal $\{\beta_i\}$ (right) overlaid by $g(\beta)$ in black; on the top row, $a = 0.5$ and $b = 7.5 \times 10^8$, on the bottom row $a = 10^{-4}$ and $b = 9.5 \times 10^3$.

2.3.1 When is the geometric temperature placement optimal?

An interesting question raised by this example is under what circumstances will tuning of the $\{\beta_i\}$ be likely to make efficiency gains over the default geometric spacing for fixed n? Suppose the target distribution is the d-dimensional multivariate Gaussian with mean μ and variance Σ. Then, $h(x) = \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$ and the tempered distributions are d-dimensional multivariate Gaussian with mean μ and variance $\beta_i^{-1}\Sigma$. More importantly, $g(\beta) = \frac{d}{2\beta}$ and $g'(\beta) = \frac{-d}{2\beta^2}$. As a result, when the $\{\beta_i\}$ are geometrically spaced, all the partial derivatives $\frac{\partial S_n}{\partial \beta_i} = 0$ in Equation (13), and so geometric spacing is also the optimal minimum S_n spacing. In fact we can go further: suppose the set of $\frac{\partial S_n}{\partial \beta_i}$ are all zero for a general g and for all n when the

		$n=2$	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$
	Optimal						
$a=0.5$	S_n	0.83386	0.30241	0.13214	0.06218	0.03023	0.01492
$b=7.5\times10^8$	α	0.78	0.80	0.84	0.87	0.91	0.93
(convex)	τ	1.55	1.48	1.38	1.28	1.20	1.14
	Geometric						
	S_n	0.90444	0.38612	0.18454	0.09122	0.04548	0.02272
	α	0.78	0.79	0.82	0.85	0.89	0.89
	τ	1.58	1.51	1.46	1.36	1.26	1.26
	Optimal						
$a=10^{-4}$	S_n	1.46627	0.63456	0.29879	0.14591	0.07234	0.03607
$b=9.5\times10^3$	α	0.55	0.63	0.72	0.80	0.85	0.90
(concave)	τ	7.05	2.36	1.75	1.47	1.33	1.22
	Geometric						
	S_n	3.34158	2.20779	1.25229	0.64996	0.32786	0.16428
	α	0.51	0.51	0.55	0.61	0.69	0.78
	τ	591.36	55.56	9.13	3.11	1.91	1.54

Table 1: Results for the Witch's Hat problem under two settings of the parameters a and b and multiple choices of n, the number of tempering levels. The minimal sum of squares, observed acceptance rates and estimated integrated autocorrelation times are shown for the geometric scheme and for the optimal scheme.

$\{\beta_i\}$ are geometrically spaced, i.e. when $\frac{\beta_{i+1}}{\beta_i} = c_n$ where $c_n = \left(\frac{\beta_n}{\beta_0}\right)^{1/n}$, $\beta_n \neq 0$, then

$$g'(\beta_i) = -\frac{g\left(\frac{\beta_i}{c_n}\right) - 2g(\beta_i) + g(c_n\beta_i)}{\frac{\beta_i}{c_n} - 2\beta_i + c_n\beta_i}, \quad i = 1, \ldots, n-1. \tag{14}$$

As $n \to \infty$ with fixed β_0 and β_n, $c_n \to 1$ and a repeated application of L'hôpital's rule yields the equation

$$\beta g'(\beta) = -(\beta^2 g''(\beta) + \beta g'(\beta)) \tag{15}$$

which has general solution $g(\beta) = \frac{K_1}{\beta} + K_2$ for constants K_1 and K_2. In other words, geometric spacing only minimises S_n if the target distribution has this form of $g(\beta)$. This is a wider class than just the Gaussian, for example the exponential distribution has $g(\beta) = \frac{1}{\beta}$. The result ties in with Figure 3 and Table 1.

2.4 An alternative perspective on the optimisation problem

In this section we provide an intuitive formalisation of what quantity S_n in Equation (7) represents. Since $p_i(x) = \pi(x)\exp(-\beta_i h(x))/Z(\beta_i)$, it follows that

$$\begin{aligned}\frac{Z(\beta_{i+1})}{Z(\beta_i)} &= \frac{p_i(x)}{\exp(-\beta_i h(x))}\frac{\exp(-\beta_{i+1}h(x))}{p_{i+1}(x)} \\ &= \exp(-(\beta_{i+1}-\beta_i)h(x))\frac{p_i(x)}{p_{i+1}(x)}\end{aligned}$$

Taking logarithms,

$$\log\left(\frac{Z(\beta_{i+1})}{Z(\beta_i)}\right) = (\beta_i - \beta_{i+1})h(x) - \log\left(\frac{p_{i+1}(x)}{p_i(x)}\right). \tag{16}$$

Multiplying both sides of the equation by $p_{i+1}(x)$ and integrating with respect to $p_{i+1}(x)$ leads to

$$\log\left(\frac{Z(\beta_{i+1})}{Z(\beta_i)}\right) = (\beta_i - \beta_{i+1})\mathsf{E}_{i+1}[h(X)] - KL\left[p_{i+1}, p_i\right], \tag{17}$$

where $KL[p_{i+1}, p_i] = \int_x p_{i+1}(x)\log\left(\frac{p_{i+1}(x)}{p_i(x)}\right)$, is the Kullback-Leibler divergence between distributions p_{i+1} and p_i. Similarly, it can be shown, by multiplying both sides of Equation (16) by $p_i(x)$ and integrating with respect to x, that

$$\log\left(\frac{Z(\beta_{i+1})}{Z(\beta_i)}\right) = (\beta_i - \beta_{i+1})\mathsf{E}_i[h(X)] + KL[p_i, p_{i+1}]. \tag{18}$$

Summing both Equations (17) and (18) over i indices leads to

$$\begin{aligned}\log\left(\frac{Z(\beta_n)}{Z(\beta_0)}\right) &= \sum_i (\beta_i - \beta_{i+1})\mathsf{E}_{i+1}[h(X)] - \sum_i KL[p_{i+1}, p_i] \\ &= \sum_i (\beta_i - \beta_{i+1})\mathsf{E}_i[h(X)] + \sum_i KL[p_i, p_{i+1}].\end{aligned}$$

It now follows directly that

$$S_n(\beta_0, \ldots, \beta_n) = \frac{1}{2}\sum_{i=0}^{n-1}\{KL[p_{i+1}, p_i] + KL[p_i, p_{i+1}]\}$$

Thus our optimisation problem can be recast as one of finding temperatures $\{\beta_1, \ldots, \beta_{n-1}\}$ to minimise the sum of the symmetrised Kullback-Leibler distances between successive distributions p_i and p_{i+1}. This interpretation ties in with the recent work by Lefebvre, Steele and Vandal (2010) who consider this same symmetrised Kullback-Leibler divergence in picking optimal schemes for path sampling. A similar perspective, but in the context of marginal likelihood estimation using the power posterior method of Friel and Pettitt (2008), appears in Section 3.2 of that paper and also in Calderhead and Girolami (2009).

3 Application of the tuning to a non-toy problem

3.1 A Bayesian mixture problem

We now turn to Bayesian mixture modelling, an application where tempered transitions has been advocated in the past as a possible solution to sampling problems (see, for example, Celeux, Hurn and Robert (2000) and Jasra, Holmes and Stephens (2005)). The benchmark for good MCMC mixing here is label-switching: In the Bayesian treatment of a $k-$component mixture model, the likelihood is invariant to the labelling of the components. This invariance is inherited by the posterior if, as is quite natural in many cases, the priors do not impose identifiability. The logical conclusion of such invariance is that a well-mixing MCMC sampler should visit all $k!$ labellings of the components. Label-switching could be achieved trivially by incorporating a move type which permutes the component labels, however this may mask more significant difficulties in moving around the state space. Certainly we can have greater confidence in the exploratory powers of a sampler which can swap component labels in the course of its other moves.

We use the much-studied galaxy data set for illustration, see for example Richardson and Green (1997), which comprises measurements on the velocities of 82 galaxies (Figure 4). Unlike in that paper though, we fix the number of mixture components at three, the smallest number of components with non-negligible posterior probability according to Richardson and Green. Using a small number of components makes label switching harder as there is no "redundant" component to move freely around the state space exchanging identities with the less mobile components needed to explain the data if they pass sufficiently close.

Denoting the 82 velocity measurements by $y = \{y_1, \ldots, y_{82}\}$, we follow Richardson and Green (1997) in incorporating corresponding latent allocation variables $z = \{z_1, \ldots, z_{82}\}$. Given $z_i = j$, y_i follows the j^{th} of the three component Gaussian distributions of the mixture,

$$f(y_i|z_i = j, \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(y_i - \mu_j)^2}{2\sigma_j^2}\right) \quad i = 1, \ldots, 82.$$

Further, conditional independence is assumed for the observations. We specify largely independent standard proper priors:

$$\begin{aligned}
\mathsf{P}(z_i = j) &= w_j, \quad \text{where } \sum_{j=1}^{3} w_j = 1 \\
\{w_1, w_2, w_3\} &\sim Dirichlet(1, 1, 1) \\
\mu_j &\sim N(0, 1000), \quad j = 1, 2, 3 \\
\sigma_j^2 &\sim InvGam(1, 1), \quad j = 1, 2, 3.
\end{aligned}$$

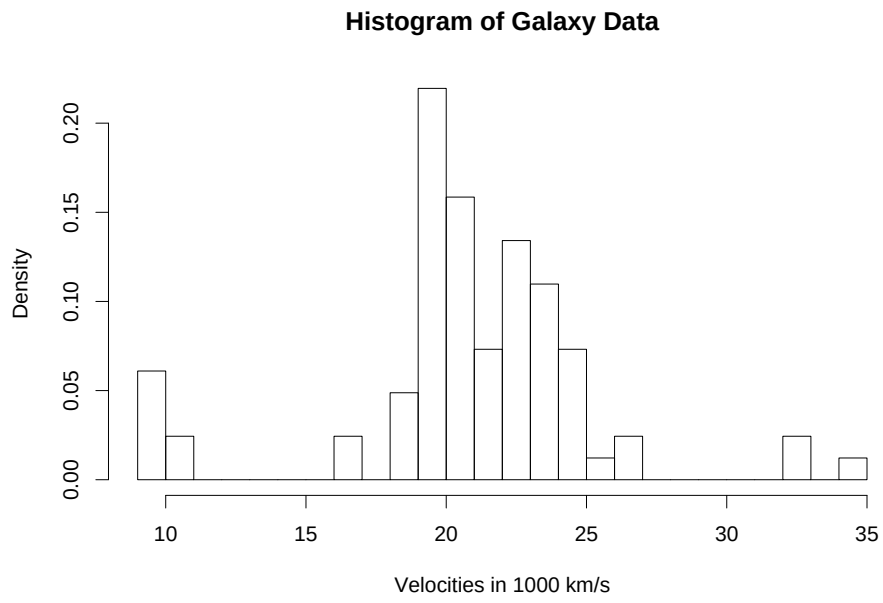

Figure 4: The galaxy data used for as illustration of the mixture modelling.

so that the posterior of interest is

$$f(z, \{w_j, \mu_j, \sigma_j^2\}_{j=1}^3 | y) \propto \prod_{i=1}^{82} f(y_i | z_i, \mu_{z_i}, \sigma_{z_i}^2) \times f(\{w_j\}) \times \prod_{j=1}^{3} f(\mu_j) \times \prod_{j=1}^{3} f(\sigma_j^2) \times \prod_{i=1}^{82} f(z_i | \{w_j\}) \quad (19)$$

We know that if label switching is taking place when sampling from Equation (19) that the marginal posterior distributions for the sets of parameters of the three Gaussian components should be identical. Figure 5 shows the output for the $\{\mu_j\}$ parameters using 100000 iterations of standard MCMC updates including a burn in of 10000 iterations (Gibbs updates for $\{w_j\}, \{\mu_j\}, \{\sigma_j^2\}$ and a uniform Metropolis proposal to change the $\{z_i\}$ in turn); it is clear that label switching is not happening. Tempering the whole of the posterior defined by Equation (19) is problematic as there is no guarantee that the tempered distributions will remain proper. Instead, we follow Celeux, Hurn and Robert (2000) in tempering only the likelihood contribution leaving the priors untempered. This approach generates proper tempered distributions provided the priors are proper. In the notation of previous sections, we set $\beta_0 = 1$ and $\beta_n = 1/16$ while

$$h(x) = \sum_{j=1}^{3} \left(\frac{n_j}{2} \log(\sigma_j^2) + \frac{1}{2\sigma_j^2} \sum_{\substack{i=1 \\ z_i = j}}^{82} (y_i - \mu_j)^2 \right)$$

where $x = \{z, \{w_j, \mu_j, \sigma_j^2\}_{j=1}^3\}$ and $n_j = \sum_{i=1}^{82} I_{[z_i=j]},\ j = 1, 2, 3$. Unlike in the motivating example of the previous section, the $g(\beta)$ corresponding to this form of $h(x)$ is not available analytically and so we must now address the question of approximating it before we can optimise the $\{\beta_i\}$.

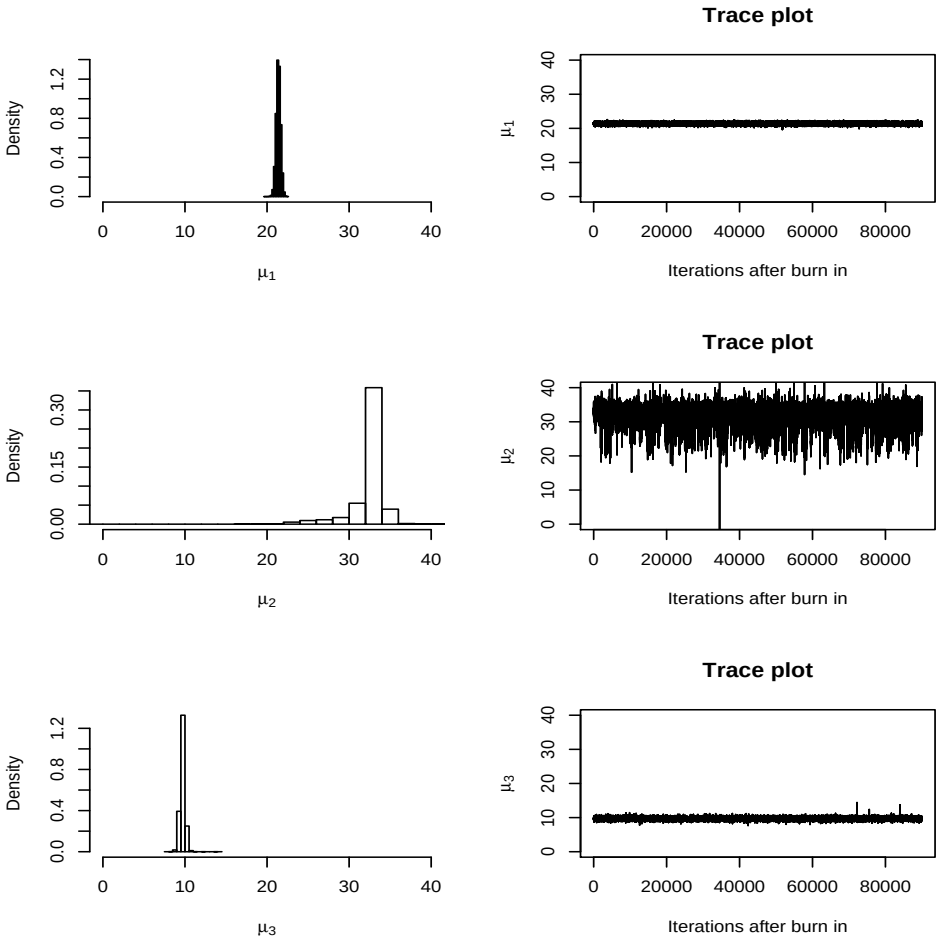

Figure 5: Histograms and trace plots of the $\{\mu_j\}$ chains indicating a lack of mode swapping when $\beta = 1$.

3.2 Approximating $g(\beta)$

The difficulty in estimating $g(\beta) = \mathsf{E}_\beta[h(X)]$ and $g'(\beta) = -(\mathsf{E}_\beta[h(X)^2] - \mathsf{E}_\beta[h(X)]^2)$ for $\beta_n \leq \beta \leq \beta_0$ is that sampling under p_β is difficult for β close to β_0 (hence the need for tempered transitions!). We propose instead to estimate $g(\beta)$ and $g'(\beta)$ using importance sampling. The obvious importance distribution to use is p_{β_n} since we have already made an assumption that we can sample from this distribution quite freely. However it may be a poor choice as an importance distribution for p_β when β is close to β_0 because when the importance distribution is quite far from the target, the resulting estimates can be dominated by a handful of the samples (Robert and Casella 1999). As a compromise, we importance sample for expectations under p_β by sampling under $p_{\tilde{\beta}}$ for some $\tilde{\beta}$ where $\beta_n \leq \tilde{\beta} < \beta \leq \beta_0$, in which case the unnormalised importance weights are $\exp(-(\tilde{\beta} - \beta)h(x))$. We note in passing that a standard result states that the importance distribution which minimises the variance of the importance estimate of some function

$\psi(x)$ is

$$f^*(x) \propto |\psi(x)|\, p_\beta(x).$$

We turn this statement around to ask for what function $\psi(x)$ is $p_{\tilde{\beta}}(x)$ the optimal importance distribution?

$$\begin{aligned} |\psi(x)| &\propto \frac{p_{\tilde{\beta}}(x)}{p_\beta(x)} \\ &= \exp(-(\tilde{\beta}-\beta)h(x)) \\ &= 1-(\tilde{\beta}-\beta)h(x)+((\tilde{\beta}-\beta)h(x))^2/2+\ldots \end{aligned}$$

So using $\tilde{p_\beta}$ as an importance distribution would be optimal if we were trying to estimate $\exp(-(\tilde{\beta}-\beta)h(x))$. It is not optimal for estimating $\mathsf{E}_\beta[h(X)]$ and $\mathsf{E}_\beta[h(X)^2]$, however it may be more reasonable for this goal when $(\tilde{\beta}-\beta)$ is small, than if we were, say, trying to estimate $\mathsf{E}_\beta[X]$ or $\mathsf{E}_\beta[X^2]$.

We work with 20 uniformly spaced values of $\tilde{\beta}$ in the interval $[\beta_n, \beta_0]$. As a compromise between the inadequate sampling for large β and the risk of unreliable importance sampling for large $\beta-\tilde{\beta}$, we generate relatively small samples at each $\tilde{\beta}$ and use these samples to estimate $g(\tilde{\beta})$ and $g'(\tilde{\beta})$ both directly and indirectly by importance sampling using the next smallest of the 20 chosen values (with obvious modifications at the end points). Figure 6 shows the results when using 10000 samples at each $\tilde{\beta}$ and discarding the first 1000 iterations as burn in. We propose to use the average of the two estimates for $g(\tilde{\beta})$ and $g'(\tilde{\beta})$ at each point, with visual inspection recommended to check for major discrepancies. In this example, the estimated $g(\beta)$ curve appears quite far from the geometric-friendly form $g(\beta)=\frac{K_1}{\beta}+K_2$.

3.3 Results

Given the importance sampling estimates of both curve $g(\beta)$ and its derivative $g'(\beta)$, we can minimise $S_n(\beta_0,\ldots,\beta_n)$ using Equation (13). As before, we use the R optimisation routine optim with linear interpolation used to evaluate g and g' between the 20 $\tilde{\beta}$ values. In order to assess the effects of the imperfect estimation of $g(\beta)$ on the tuning procedure, we replicate the estimation process five times with each replicate being used to select $\{\beta_i\}$. Figure 7 shows both the variability in estimated g and g' and how the optimised S_n decreases with n for the five sets of estimates. By letting n become sufficiently large, it would be possible to reduce S_n below any positive threshold. (An upper bound on the minimum S_n is $\frac{1}{n}(\beta_0-\beta_n)(g(\beta_n)-g(\beta_0))$, achieved either by uniformly spacing $\beta_1,\ldots,\beta_{n-1}$ or by uniformly spacing $g(\beta_1),\ldots,g(\beta_{n-1})$.) However as the computational cost of the tempering increases linearly in n, the curves show that costs grow quite rapidly for relatively small decreases in S_n.

Turning to the tempered transitions themselves, we ran the algorithm for 100000 iterations including a burn-in of 10000 iterations. At each tempering stage, the same proposal types were used as in the importance

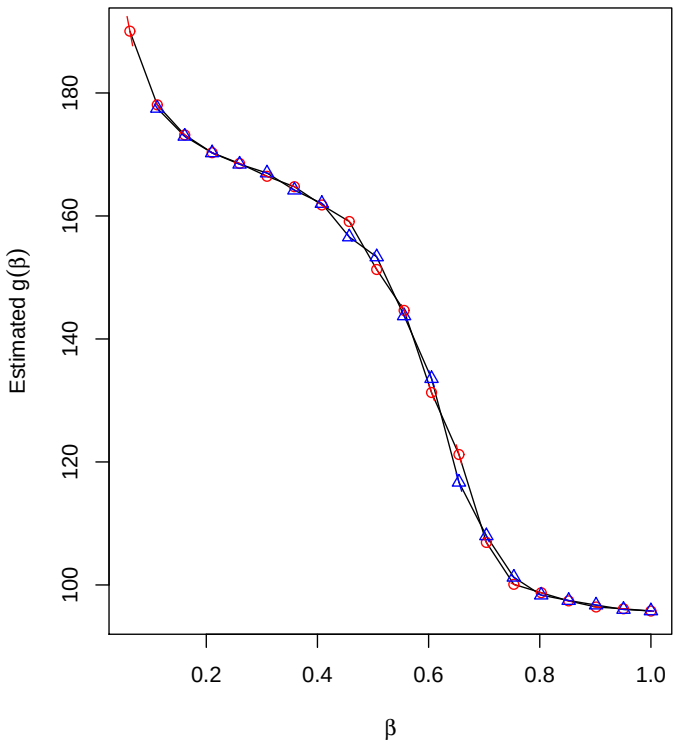

Figure 6: Estimating $g(\beta)$ and $g'(\beta)$: Symbols indicate the $\tilde{\beta}$ at which samples are generated; red circles and red lines segments indicate direct sampling; blue triangles and blue line segments indicate importance sampling estimates; black lines are linear interpolations.

sampling. For each of the five sets of importance estimates, we temper using $n = 64, 128, 256, 512$ and compare the optimised β results with those of geometric spacing. We know that if switching is taking place, then the marginals for each component should be identical. This obviously implies that the three posterior expected μ_i should be equal, as should the expected weights and the expected variances. We propose to monitor the mixing using the usual tool of the integrated autocorrelation times, however in estimating these diagnostics we use the averages of each group of parameters over the three chains rather than the usual chain-wise average for each parameter. For example, if the labels swap regularly, the individual averages of each of the three μ_i chains will be close to their overall average, and the non-centred autocorrelations will not be much different from the standard centred ones. On the other hand, if the labels do not switch as was the case with standard MCMC illustrated in Figure 5, the autocorrelations calculated around the overall average will be greatly increased and this will be reflected in the modified integrated autocorrelation times.

Table 2 summarises the results, showing the acceptance rates and the estimated integrated autocorrela-

	$n = 64$	$n = 128$	$n = 256$	$n = 512$
Geometric	α=0.00013	α=0.00065	α=0.00187	α=0.00853
$\hat{\tau}(\{w_j\})$	***, ***, ***	3253, 22910, 95995	1649, 1382, 1916	256, 249, 257
$\hat{\tau}(\{\mu_j\})$	96042, ***, ***	4505, 6938, 4552	1003, 1457, 745	322, 322, 269
$\hat{\tau}(\{\sigma_j^2\})$	***, 4970 5188	477, 768, 1137	710, 475, 799	140, 132, 163
Tuned 1	α=0.00062	α=0.00316	α=0.01366	α=0.04105
$\hat{\tau}(\{w_j\})$	9089, 4750, 5743	1137, 721, 1117	161, 148, 161	50, 51, 45
$\hat{\tau}(\{\mu_j\})$	6406, 3842, 2702	781, 762, 952	134, 165, 160	45, 48, 49
$\hat{\tau}(\{\sigma_j^2\})$	3071, 1323, 5498	321, 113, 441	78, 88, 90	31, 33, 28
Tuned 2	α=0.00054	α=0.00346	α=0.01426	α=0.04194
$\hat{\tau}(\{w_j\})$	5481, 10701, 17267	642, 1357, 952	134, 161, 148	51, 45, 49
$\hat{\tau}(\{\mu_j\})$	13077, 6291, 20823	893, 1235, 1309	148, 123, 133	48, 54, 51
$\hat{\tau}(\{\sigma_j^2\})$	6040, 8845, 2274	249, 250, 239	116, 72, 44	38, 34, 26
Tuned 3	α=0.00053	α=0.00362	α=0.01395	α=0.04467
$\hat{\tau}(\{w_j\})$	8610, 4050, 3609	774, 1244, 813	185, 166, 178	49, 50, 49
$\hat{\tau}(\{\mu_j\})$	6826, 19664, 25232	1499, 827, 1133	185, 167, 149	46, 47, 51
$\hat{\tau}(\{\sigma_j^2\})$	5967, 2569, 2283	268, 288, 156	83, 91, 56	34, 37, 23
Tuned 4	α=0.00054	α=0.00282	α=0.00923	α=0.03884
$\hat{\tau}(\{w_j\})$	4668, 8875, 8538	960, 1043, 862	258, 253, 244	51, 50, 55
$\hat{\tau}(\{\mu_j\})$	6335, 3805, 16416	2365, 751, 1928	250, 197, 209	51, 47, 53
$\hat{\tau}(\{\sigma_j^2\})$	3266, 4463, 1782	411, 218, 376	91, 115, 117	43, 32, 39
Tuned 5	α=0.00062	α=0.00275	α=0.00928	α=0.02518
$\hat{\tau}(\{w_j\})$	18057, 6343, 7629	1065, 763, 1353	293, 261, 210	75, 77, 86
$\hat{\tau}(\{\mu_j\})$	5513, 5965, 39982	3743, 1671, 1966	221, 272, 287	91, 93, 87
$\hat{\tau}(\{\sigma_j^2\})$	1423, 1496, 538	492, 506, 386	118, 149, 118	59, 48, 64

Table 2: Results for the mixture problem using different numbers of tempering levels. Results are shown for geometric spacing of the $\{\beta_i\}$ and for the tuned spacing from the five replicates of importance sampling. *** indicates that the estimates of integrated autocorrelation times did not converge reliably.

	$n = 64$	$n = 128$	$n = 256$	$n = 512$
Estimating g and g'	8.35	8.35	8.35	8.35
Optimisation	2.78	16.97	52.02	191.00
Tuned tempering	879.81	1762.61	3507.79	7022.67
Tuned total	890.94	1787.93	3568.16	7222.02
Geometric tempering	874.12	1758.88	3494.94	7002.63
Proportion increase	1.019	1.017	1.021	1.031

Table 3: Time in user CPU time seconds for the geometrically spaced temperatures and for the tuned temperatures including a breakdown of the cost of tuning.

each of only 10000 iterations, for the $g(\beta)$ and $g'(\beta)$ estimation; this stage is independent of n. The cost of the deterministic minimisation of S_n in R increases with n but is of the order of a few minutes for $n = 512$. This makes the cost of the tuning procedure a small fraction of the total cost. Full details are given in Table 3 showing that the additional cost of the tuning procedure, for this example, varied between 2% and 4% extra CPU time compared to the untuned procedure. Combining this information with the integrated autocorrelation times in Table 2 indicates the tuned procedure gives substantial improvements in mixing compared to the geometric temperature placement.

In this example, very little of the total computational effort was spent in estimating $g(\beta)$ and $g'(\beta)$. Although importance sampling cannot be guaranteed to be particularly good for this type of problem, we suggest that this is a sensible strategy. The associated risk is either that the importance sampling fails to identify a $g(\beta)$ curve which is not suitable for geometric spacing or, conversely, that it identifies interesting features which are not in fact present. In the former case, a visual inspection of the roughly estimated $g(\beta)$ may suggest that it is not worthwhile implementing any optimisation, reverting to the default geometric, and so the wasted CPU time is minimal. (The same argument is also reasonable when importance sampling works well for estimating $g(\beta)$.) On the other hand, if the estimated $g(\beta)$ looks to be of the form where tuning may help, more computational effort could be put into improving the accuracy of its estimation, especially if a discrepancy is noted between the values of the curve using direct and indirect sampling.

4 Discussion

In this paper we have explored how to tune the expensive tempered transitions algorithm to make best use of computational resources. We have shown that the geometric schedule will be optimal if the curve

$g(\beta) = \mathsf{E}_\beta[h(X)]$ is of a particular form, where the target distribution is $p(x) \propto \pi(x)\, \exp(-\beta_0\, h(x))$. The tuning itself is relatively cheap and examples have demonstrated that it can make a significant difference.

Although we have not explicitly considered the question of choosing the number of tempering levels, we have some purely anecdotal evidence that the tuning procedure may yield useful information regarding the minimum number of tempering levels required. In our experience, tempered transitions does not seem to perform at all well with a $\{\beta_i\}$ for which $S_n > 1$. For example, in our mixture example the geometric S_n is approximately of the order 2, 1, 0.5 and 0.25 for $n = 64, 128, 256, 512$ respectively, while for the tuned $\{\beta_i\}$ it is of the corresponding approximate value 1.2, 0.6, 0.3 and 0.15. It also seems feasible that the tuning approach proposed here may also be relevant to some of the other MCMC algorithms which incorporate an element of tempering. This is another topic for future research.

Acknowledgements

Gundula Behrens thanks the Engineering and Physical Sciences Research Council and Evangelisches Studienwerk for financial support. Nial Friel's research was supported by a visit to Bath by the Bath Institute for Complex Systems (EPSRC grant GR/S86525/01) and by a Science Foundation Ireland Research Frontiers Program grant, 09/RFP/MTH2199. We are grateful to the Associate Editor and the referees for their insightful and helpful comments and to Dr Jey Sivaloganathan for advice on Equations (14) and (15).

References

[1] B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, 53(12):4028–4045, 2009.

[2] G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

[3] N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, Series B*, 70:589–607, 2008.

[4] A. Gelman and X-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

[5] C.J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics*, 23:156–163, 1991.

[6] C.J. Geyer and E.A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[7] R. Gramacy, R. Samworth, and R. King. Importance tempering. *Statistics and Computing*, 20:1–7, 2010.

[8] Y. Iba. Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, 12(5):623–656, 2001.

[9] A. Jasra, C.C. Holmes, and D.A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science*, 17(1):50–67, 2005.

[10] A. Jasra, D.A. Stephens, and C.C. Holmes. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.

[11] G. Lefebvre, R.J. Steele, and A.C. Vandal. A path sampling identity for computing the Kullback-Leibler and J-divergences. *Computational Statistics and Data Analysis*, 54(7):1719–1731, 2010.

[12] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458, 1992.

[13] P. Matthews. A slowly mixing Markov chain with implications for Gibbs sampling. *Statistics and Probability Letters*, 17:231–236, 1993.

[14] R.M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

[15] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[16] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, 59(4):731–792, 1997.

[17] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, New York, 1999.