Technical Report No. 9421, Department of Statistics, University of Toronto

Sampling from Multimodal Distributions Using Tempered Transitions

Radford M. Neal

Dept. of Statistics and Dept. of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 1A1
radford@stat.toronto.edu

29 October 1994

Abstract. I present a new Markov chain sampling method appropriate for distributions with isolated modes. Like the recently-developed method of "simulated tempering", the "tempered transition" method uses a series of distributions that interpolate between the distribution of interest and a distribution for which sampling is easier. The new method has the advantage that it does not require approximate values for the normalizing constants of these distributions, which are needed for simulated tempering, and can be tedious to estimate. Simulated tempering performs a random walk along the series of distributions used. In contrast, the tempered transitions of the new method move systematically from the desired distribution, to the easily-sampled distribution, and back to the desired distribution. This systematic movement avoids the inefficiency of a random walk, an advantage that unfortunately is cancelled by an increase in the number of interpolating distributions required. Because of this, the sampling efficiency of the tempered transition method in simple problems is similar to that of simulated tempering. On more complex distributions, however, simulated tempering and tempered transitions may perform differently. Which is better depends on the ways in which the interpolating distributions are "deceptive".

1 Introduction

Monte Carlo methods based on sampling using Markov chains have long been used in statistical physics, and have recently become popular in Bayesian inference (for reviews, see Kennedy 1990, Smith and Roberts 1993, Neal 1993). These methods encounter problems when the distribution to be sampled has isolated modes, which the Markov chain moves between only rarely. In this case, convergence of the Markov chain to the desired distribution will be slow, and there will be large dependencies between successive states.

One approach to sampling from such a multimodal distribution is to run the sampling procedure several times, with different random number seeds, in the hopes that different runs will end up in different modes. Simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983) is sometimes used within each run in an attempt to avoid finding modes with low total probability. Simulated annealing samples (approximately) from a series of distributions, each only slightly different from the last, starting with a distribution that is easily sampled from, and ending in the distribution of interest. The last state drawn from each distribution is used to initialize the Markov chain used for sampling from the next distribution, in the hope that such initial states will be good starting points, speeding the convergence of the Markov chains. Often, a series of "canonical" distributions is used, obtained by varying a "temperature" parameter, starting at a high temperature, and cooling to a lower final temperature that produces the desired distribution. (In many applications of simulated annealing, the final temperature is zero, and the result is a solution to an optimization problem. Here, however, I am interested in sampling from non-degenerate distributions at a non-zero final temperature.)

Unfortunately, multiple independent runs will not in general produce a sample in which each mode is fairly represented, because the probability of a run reaching a mode will depend more on the mode's "basin of attraction" than on the total probability in the vicinity of the mode. This remains true even when simulated annealing is used. Annealing may, however, reduce the magnitude of this discrepancy, provided the distribution at a high temperature is a good guide to the distribution at the lower final temperature.

In contrast, the method of simulated tempering (Marinari and Parisi 1992, Geyer and Thompson 1994) is guaranteed to sample from the correct distribution, given enough time. Like annealing, it uses a series of distributions (often parameterized by temperature), but instead of proceeding through these distributions (varying the temperature) systematically, simulated tempering treats the index of the distribution as an additional variable to be updated stochastically. Different temperatures are therefore explored via a random walk. At times when the temperature is set to its lowest value, the rest of the state will have the desired distribution. At other times, the temperature will be high, and it will be easy to move from the region of one mode to that of another. Thus a Markov chain for exploring this extended state space may be able to move easily from one mode to another even though these modes are isolated in the original state space.

To use simulated tempering, we must define a joint distribution for the temperature together with the original state variables, in which the distribution for the original state variables conditional on the temperature taking its lowest value is the distribution of interest. This joint distribution must also lead to a marginal distribution for the temperature that is approximately uniform, so that the Markov chain will be able to move from low to high temperature and back. Defining such a joint distribution requires approximate estimates for the normalizing constants of the distributions over the original state variables at the various temperatures used. These normalizing constants will generally be estimated using preliminary runs, a process that can be tedious and time-consuming.

Two other methods have also been proposed for sampling from multimodal distributions.

The multicanonical method (Berg and Celik 1992) can be viewed as an elaboration of simulated tempering in which the temperature parameter is eliminated by explicit summation. Its properties are similar to those of simulated tempering; in particular, it too explores via a random walk, and requires preliminary estimates for normalization constants. Metropolis-coupled Markov chains (Geyer 1991) maintain state variables for each interpolating distribution (temperature). These states are updated using both transitions operating on each state independently, and transitions that may swap states at adjacent temperatures. This method does not require estimates for normalization constants, but the maintenance of many states is costly in both time and space.

In this paper, I describe a new sampling method based on "tempered transitions". Such a transition starts with the current state, which will have the low-temperature distribution in which we are interested, and proceeds from there through a series of states sampled (approximately) from the various interpolating distributions, with the temperature increasing in the first half of the series, and decreasing in the second half. The state at the end of this series is then accepted or rejected based on ratios of probabilities computed during the intermediate stages.

The tempered transition method has the major advantage that it does not require preliminary estimates for normalization constants. It also avoids the inefficiency of moving between interpolating distributions via a random walk, but this benefit turns out to be cancelled by the need for a larger number of interpolating distributions, with the result that in simple problems the sampling efficiency of tempered transitions is similar to that of simulated tempering. In more complex problems, the two methods can perform differently, however. Which is better depends on the ways in which the interpolating distributions are "deceptive".

In Section 2, which follows, I review the simulated tempering method. In Section 3, I present the new method of tempered transitions, show that it leads to a Markov chain for which the desired distribution is invariant, and discuss how the method operates when applied to canonical distributions controlled by a temperature parameter. Section 4 examines how simulated tempering and tempered transitions perform when sampling from a multivariate Gaussian, with the conclusion that the two methods perform similarly in this simple case. Section 5 demonstrates the methods on a more complex test problem, and discusses how in this case the deceptive aspects of the interpolating distributions can affect which method performs best. Finally, Section 6 discusses the advantages and disadvantages of existing tempering methods, and the prospects for developing new methods that combine the advantages of several existing methods.

2 Simulated tempering

Suppose that we wish to sample from a distribution given by the probability or probability density function $p_0(x)$, which may have many isolated modes, typically in a high-dimensional space. A series of n other distributions, $p_1(x), \ldots, p_n(x)$, are also defined, with each p_i being similar to p_{i-1}. The end-point of the sequence, p_n, is quite different from p_0, however, and is thought to be easier to sample from. The p_i will generally be available only in unnormalized form, with $p_i(x) = f_i(x)/Z_i$, where the $f_i(x)$ are known functions. The normalization constants, $Z_i = \int f_i(x)\,dx$, are unknown, at least initially.

Often, the f_i will be defined in terms of a common "energy" function, $E(x)$, and a series of "temperatures", T_i, or inverse temperatures, $\beta_i = 1/T_i$, with $f_i(x) = \exp(-\beta_i E(x))$. This yields corresponding "canonical" probability distributions:

$$p_i(x) \;=\; \frac{1}{Z_i}\exp\left(-\beta_i E(x)\right) \tag{1}$$

Depending on the problem, it may be appropriate to define a series of interpolating distributions in some other fashion, guided by whatever knowledge we may have of the properties of the distribution of interest, but in this paper the above form will be assumed whenever a specific form is required for the discussion.

In the simulated tempering method (Marinari and Parisi 1992, Geyer and Thompson 1994), a Markov chain is defined whose states consist of both a value for x and an index, i, of one of the distributions in the series. This Markov chain is set up to converge to the equilibrium distribution given by:

$$p(i,x) \;\propto\; w_i f_i(x) \tag{2}$$

for some constants w_i, which I will refer to as the weights for the distributions. Note that the conditional distribution for x given some value of i is $p_i(x)$. A sample of (dependent) points from p_0 can thus be obtained by simulating the Markov chain for some period of time and then discarding all but those states for which $i = 0$.

Markov chains used for simulated tempering employ transitions of two types, which will usually be applied alternately. In transitions of the first type, which I will call "base transitions", i is kept fixed, while x is updated in some fashion that leaves $p_i(x)$ invariant. This update of x might be done using Gibbs sampling for the components of x, or using some form of the Metropolis algorithm, or using whatever other method seems appropriate for the particular problem. Transitions of the second type will leave x unchanged, but may change i. This can be done in several ways, between which there is little to choose. I will assume here that i is updated by the Metropolis algorithm, using a proposal distribution in which $i' = i+1$ and $i' = i-1$ are equally likely. The proposal is rejected if i' is outside the valid range (0 to n); otherwise it is accepted with probability $\min[1, p(i',x)/p(i,x)]$, which for canonical distributions is

$$\min\left[1,\ \frac{w_{i'}}{w_i}\exp((\beta_i - \beta_{i'})E(x))\right] \tag{3}$$

If the distributions in the series are all to play a useful role in sampling — in particular, if the high-temperature distributions are to facilitate movement between modes — it is necessary that the Markov chain spend roughly equal amounts of time at each value of i. The marginal distribution of i with respect to the equilibrium distribution is given by

$$p(i) \;=\; \int p(i,x)\,dx \;\;\propto\;\; \int w_i f_i(x)\,dx \;=\; w_i Z_i \tag{4}$$

A roughly uniform distribution over i can therefore be obtained by letting each weight, w_i, be approximately $1/Z_i$, or some constant multiple thereof. Note that the Z_i will typically vary over many orders of magnitude.

Since the Z_i are initially unknown, suitable values for the weights must be found through a process of trial and error, using preliminary runs. To do this, we simulate the Markov chain using the current values for the w_i, and observe the resulting frequencies, q_i, with which each distribution is visited. New, better weights, w'_i are then calculated as $w'_i = w_i/q_i$. In the initial stages, some of the observed frequencies q_i may be zero, in which case some extrapolation is necessary. Various other elaborations of the estimation procedure are also possible (Geyer and Thompson 1994).

This process of finding suitable weights for the distributions used in simulated tempering can be tedious and time-consuming. The major advantage of the tempered transition method to be described next is that it does not require that preliminary runs be performed in order to find such weights. It should be noted, however, that both methods may require preliminary runs in order to select an appropriate number, n, of interpolating distributions, and to determine how the temperatures, T_i, for these distributions should be spaced.

3 The tempered transition method

Suppose, as above, that we wish to sample from a distribution given by $p_0(x)$, which may have many isolated modes, and that to assist in this, a series of n other distributions, $p_1(x), \ldots, p_n(x)$, are also defined, with p_n being easier to sample from than p_0. I will here describe the new "tempered transition" method for utilizing this series of distributions in sampling.

3.1 Applying the tempered transition method

To use tempered transitions, we must have, for each i, a pair of base transitions, $\hat{T}_i$ and $\check{T}_i$, which both have p_i as an invariant distribution, and which satisfy the following mutual reversibility condition for all x and x':

$$p_i(x)\,\hat{T}_i(x,x') \;=\; \check{T}_i(x',x)\,p_i(x') \tag{5}$$

$\hat{T}_i$ and $\check{T}_i$ may be identical, in which case it must satisfy detailed balance with respect to p_i. If $\hat{T}_i$ is composed of several sub-transitions applied in sequence, $\hat{T}_i = S_1 \cdots S_k$, and the S_j all satisfy detailed balance, then we can let $\check{T}_i = S_k \cdots S_1$.

A tempered transition first finds a candidate state by applying the base transitions in the sequence $\hat{T}_1 \cdots \hat{T}_n \check{T}_n \cdots \check{T}_1$. This candidate state is then accepted or rejected based on ratios of probabilities involving intermediate states. Since the generation of the candidate state involves use of $\hat{T}_n$ and $\check{T}_n$, which are presumed to move about the state space rapidly, we may hope that the candidate state will have a wide distribution, not confined to the mode in which the start state is located. The principal role of the intermediate transitions is to keep the probability of acceptance reasonably high.

In detail, if we are currently in state $\hat{x}_0$, the candidate state, $\check{x}_0$, is generated as follows:

$$\begin{array}{c}
\text{Generate } \hat{x}_1 \text{ from } \hat{x}_0 \text{ using } \hat{T}_1. \\
\text{Generate } \hat{x}_2 \text{ from } \hat{x}_1 \text{ using } \hat{T}_2. \\
\vdots \\
\text{Generate } \bar{x}_n \text{ from } \hat{x}_{n-1} \text{ using } \hat{T}_n. \\
\text{Generate } \check{x}_{n-1} \text{ from } \bar{x}_n \text{ using } \check{T}_n. \\
\vdots \\
\text{Generate } \check{x}_1 \text{ from } \check{x}_2 \text{ using } \check{T}_2. \\
\text{Generate } \check{x}_0 \text{ from } \check{x}_1 \text{ using } \check{T}_1.
\end{array}$$

The candidate state is then accepted with probability

$$\min\left[1,\ \frac{p_1(\hat{x}_0)}{p_0(\hat{x}_0)} \cdots \frac{p_n(\hat{x}_{n-1})}{p_{n-1}(\hat{x}_{n-1})} \cdot \frac{p_{n-1}(\check{x}_{n-1})}{p_n(\check{x}_{n-1})} \cdots \frac{p_0(\check{x}_0)}{p_1(\check{x}_0)}\right] \tag{6}$$

If the candidate state is not accepted, the next state of the Markov chain is the same as the old state. Note that each p_i occurs an equal number of times in the numerator and denominator of the above product of ratios. The acceptance probability can therefore be computed without knowledge of the normalization constants for these distributions.

If the acceptance probability is to be reasonably high, properly-spaced intermediate distributions will have to be provided to interpolate from p_0 to p_n. Deciding on the number and spacing of such distributions may require that preliminary runs be performed. The effort in this respect should be less than that required for the simulated tempering and multicanonical approaches, where weighting factors must be found as well.

3.2 Proof of detailed balance for tempered transitions

To show that the tempered transitions described above leave the distribution p_0 invariant, it suffices to show that the detailed balance condition is satisfied for accepted transitions. Any transition from $\hat{x}_0$ to $\check{x}_0$ will involve some sequence of intermediate states, $\hat{x}_1, \ldots, \hat{x}_{n-1}, \bar{x}_n, \check{x}_{n-1}, \ldots, \check{x}_1$. I will show that the probability of starting from $\hat{x}_0$, going through these intermediate states, and accepting the end-point, $\check{x}_0$, is the same as the probability of starting in $\check{x}_0$, going through the same sequence of intermediate states, but in reverse order, and accepting the end-point, $\hat{x}_0$. From this, detailed balance for the tempered transitions follows.

First, consider the probability that we will start in $\hat{x}_0$, proceed through intermediate states $\hat{x}_1, \ldots, \hat{x}_{n-1}, \bar{x}_n, \check{x}_{n-1}, \ldots, \check{x}_1$, and finally produce $\check{x}_0$ as the candidate state. This probability is as follows (using $\hat{x}_n$ and $\check{x}_n$ as synonyms for $\bar{x}_n$):

$$\begin{aligned}
& p_0(\hat{x}_0)\,\hat{T}_1(\hat{x}_0,\hat{x}_1)\cdots\hat{T}_n(\hat{x}_{n-1},\bar{x}_n)\,\check{T}_n(\bar{x}_n,\check{x}_{n-1})\cdots\check{T}_1(\check{x}_1,\check{x}_0) \\
&= p_0(\hat{x}_0)\left[\prod_{i=1}^{n}\hat{T}_i(\hat{x}_{i-1},\hat{x}_i)\right]\cdot\left[\prod_{i=1}^{n}\check{T}_i(\check{x}_i,\check{x}_{i-1})\right] && (7) \\
&= p_0(\hat{x}_0)\left[\prod_{i=1}^{n}\frac{p_i(\hat{x}_{i-1})}{p_i(\hat{x}_{i-1})}\hat{T}_i(\hat{x}_{i-1},\hat{x}_i)\right]\cdot\left[\prod_{i=1}^{n}\frac{p_i(\check{x}_i)}{p_i(\check{x}_i)}\check{T}_i(\check{x}_i,\check{x}_{i-1})\right] && (8) \\
&= p_0(\hat{x}_0)\left[\prod_{i=1}^{n}\frac{p_i(\hat{x}_i)}{p_i(\hat{x}_{i-1})}\check{T}_i(\hat{x}_i,\hat{x}_{i-1})\right]\cdot\left[\prod_{i=1}^{n}\frac{p_i(\check{x}_{i-1})}{p_i(\check{x}_i)}\hat{T}_i(\check{x}_{i-1},\check{x}_i)\right] && (9) \\
&= \left[\prod_{i=1}^{n}\frac{p_{i-1}(\hat{x}_{i-1})}{p_i(\hat{x}_{i-1})}\check{T}_i(\hat{x}_i,\hat{x}_{i-1})\right]p_n(\bar{x}_n)\cdot\frac{1}{p_n(\bar{x}_n)}\left[\prod_{i=1}^{n}\frac{p_i(\check{x}_{i-1})}{p_{i-1}(\check{x}_{i-1})}\hat{T}_i(\check{x}_{i-1},\check{x}_i)\right]p_0(\check{x}_0) && (10) \\
&= \frac{p_0(\hat{x}_0)}{p_1(\hat{x}_0)}\cdots\frac{p_{n-1}(\hat{x}_{n-1})}{p_n(\hat{x}_{n-1})}\cdot\frac{p_n(\check{x}_{n-1})}{p_{n-1}(\check{x}_{n-1})}\cdots\frac{p_1(\check{x}_0)}{p_0(\check{x}_0)} \\
&\qquad p_0(\check{x}_0)\,\hat{T}_1(\check{x}_0,\check{x}_1)\cdots\hat{T}_n(\check{x}_{n-1},\bar{x}_n)\,\check{T}_n(\bar{x}_n,\hat{x}_{n-1})\cdots\check{T}_1(\hat{x}_1,\hat{x}_0) && (11)
\end{aligned}$$

Multiplying this by the acceptance probability (6), we see that the probability of a transition from $\hat{x}_0$ to $\check{x}_0$ passing through this sequence of intermediate states is

$$\begin{aligned}
& p_0(\check{x}_0)\,\hat{T}_1(\check{x}_0,\check{x}_1)\cdots\hat{T}_n(\check{x}_{n-1},\bar{x}_n)\,\check{T}_n(\bar{x}_n,\hat{x}_{n-1})\cdots\check{T}_1(\hat{x}_1,\hat{x}_0) \\
&\quad\times\ \min\left[1,\ \frac{p_1(\check{x}_0)}{p_0(\check{x}_0)}\cdots\frac{p_n(\check{x}_{n-1})}{p_{n-1}(\check{x}_{n-1})}\cdot\frac{p_{n-1}(\hat{x}_{n-1})}{p_n(\hat{x}_{n-1})}\cdots\frac{p_0(\hat{x}_0)}{p_1(\hat{x}_0)}\right] && (12)
\end{aligned}$$

which is the same as the probability of a transition from $\check{x}_0$ to $\hat{x}_0$ through the reversed sequence of intermediate states. Since reversal of state sequences is a one-to-one mapping, this implies that detailed balance holds for tempered transitions, which therefore leave the desired distribution, p_0, invariant.

3.3 Tempered transitions for canonical distributions

To obtain an intuitive picture of how tempered transitions operate, we can look at the case of sampling from a "canonical" distribution given by $p_0(x) \propto \exp(-\beta_0 E(x))$, where E is an "energy" function, and β_0 is an inverse "temperature". We define the other distributions by $p_i(x) \propto \exp(-\beta_i E(x))$, for some β_i such that $\beta_n < \cdots < \beta_1 < \beta_0$.

In the first half of a tempered transition for this system the temperature rises (i.e. the inverse temperature falls); in the second half, the temperature does down again. The second half is similar to a simulated annealing run. If we did several simulated annealing runs starting from some initial distribution, we would have no rigorous way of choosing between them, but using tempered transitions we can decide whether to accept or reject the result of the annealing process in a valid manner.

The acceptance probability (6) for this system can be rewritten as:

$$\min \left[1,\ \frac{p_1(\hat{x}_0)}{p_0(\hat{x}_0)} \cdots \frac{p_n(\hat{x}_{n-1})}{p_{n-1}(\hat{x}_{n-1})} \cdot \frac{p_{n-1}(\check{x}_{n-1})}{p_n(\check{x}_{n-1})} \cdots \frac{p_0(\check{x}_0)}{p_1(\check{x}_0)}\right] \;=\; \min\left[1,\ \exp\left(-(\check{F} - \hat{F})\right)\right] \tag{13}$$

where $\hat{F}$ and $\check{F}$ are defined as follows:

$$\hat{F} \;=\; \log\left[\prod_{i=0}^{n-1} \frac{p_{i+1}(\hat{x}_i)}{p_i(\hat{x}_i)}\right] \;=\; \sum_{i=0}^{n-1} (\beta_i - \beta_{i+1})\, E(\hat{x}_i) \tag{14}$$

$$\check{F} \;=\; \log\left[\prod_{i=0}^{n-1} \frac{p_{i+1}(\check{x}_i)}{p_i(\check{x}_i)}\right] \;=\; \sum_{i=0}^{n-1} (\beta_i - \beta_{i+1})\, E(\check{x}_i) \tag{15}$$

$\hat{F}$ and $\check{F}$ are both "thermodynamic integration" estimates of $F = \log Z(\beta_n) - \log Z(\beta_0)$, where $Z(\beta) = \int \exp(-\beta E(x))\, dx$ is the "partition function" for the system. This is seen as follows:

$$F \;=\; -\int_{\beta_n}^{\beta_0} \frac{d}{d\beta} \log Z(\beta)\, d\beta \;=\; -\int_{\beta_n}^{\beta_0} \frac{Z'(\beta)}{Z(\beta)}\, d\beta \tag{16}$$

$$\;=\; -\int_{\beta_n}^{\beta_0} \left[\int -E(x)\, \exp(-\beta E(x))\, /\, Z(\beta)\, dx\right] d\beta \tag{17}$$

$$\;=\; \int_{\beta_n}^{\beta_0} E_\beta\, d\beta \tag{18}$$

where, E_β is the expected value of E with respect to the canonical distribution at inverse temperature β.

$\check{F}$ and $\hat{F}$ will approximate F increasingly well as the spacing of the β_i goes to zero, provided the distributions of $\hat{x}_i$ and $\check{x}_i$ approach the canonical distribution at inverse temperature β_i. The acceptance probability for the tempered transitions should therefore approach one. In practice, the presence of multiple modes may inhibit convergence to the true distribution, though there may be local convergence to the distribution in the vicinity of one mode. In such cases, the difference between $\hat{F}$ and $\check{F}$ will incorporate information on whether the mode reached by a tempered transition has more or less probability mass in its vicinity than the current mode.

The relationship of the acceptance probability to the integral F is illustrated in Figure 1. These pictures ignore fluctuations in E, showing only its expected value with respect to realizations of the sequence of transitions applied. In the top picture, it is assumed that for each i, the base transitions $\hat{T}_i$ and $\check{T}_i$ both produce a state distributed according to the distribution p_i, regardless of the previous state, so that equilibrium is established immediately at every step of the tempered transition. In this situation, the expected magnitude of $\check{F} - \hat{F}$, which determines the acceptance probability of equation (13), is the difference between an approximation to the integral F based on a step function lying above the true curve of expected energy, and the corresponding approximation based on a step function lying below the true curve. The magnitude of this difference between approximations is controlled by the number and spacing of the β_i, which must be chosen to give a reasonable acceptance probability.

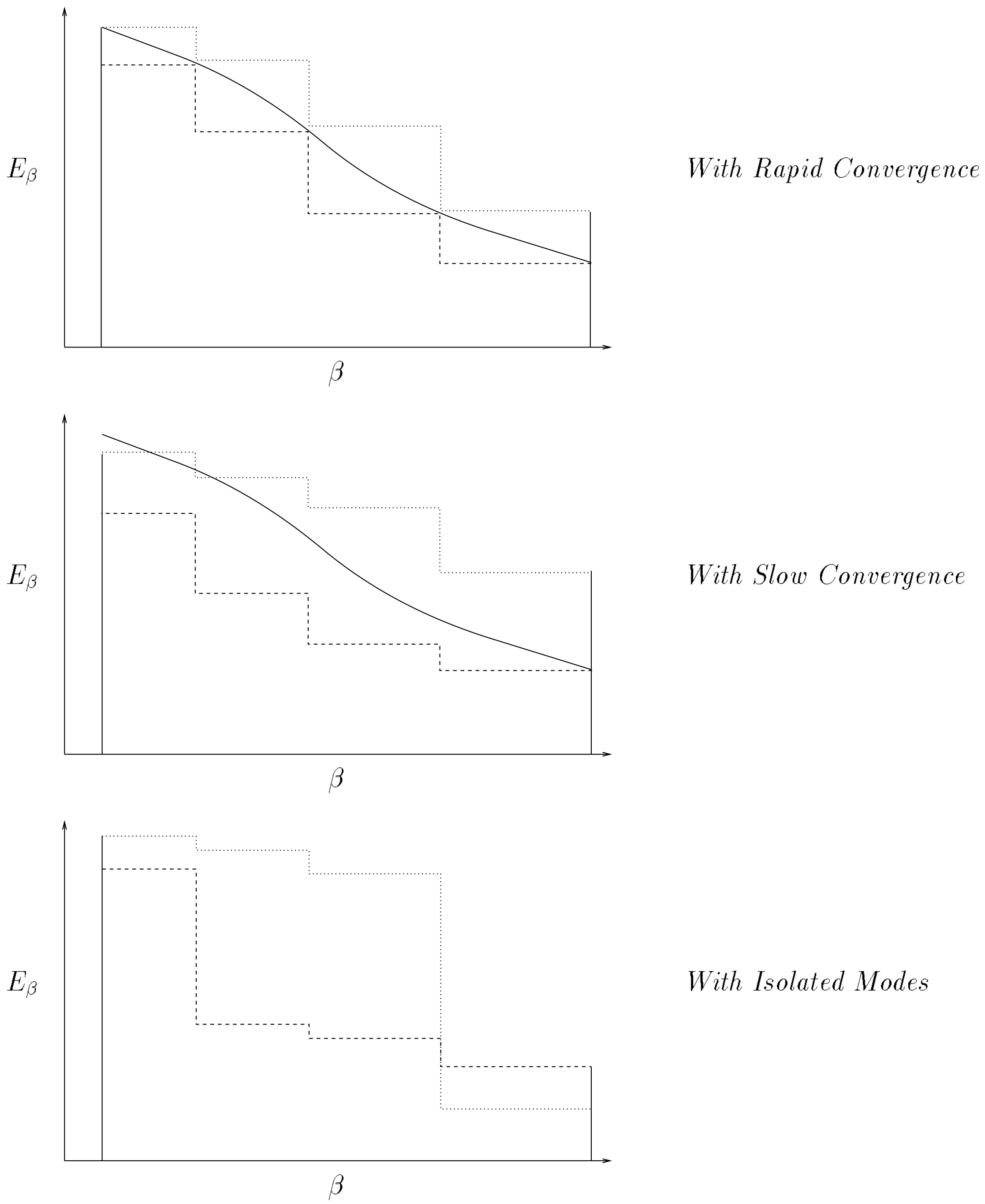

Figure 1: Relation of acceptance probability to integral approximation. These graphs show the integral approximations on which the acceptance probability for a tempered transition depends, for situations of rapid convergence, slow convergence, and convergence strongly inhibited by isolated modes. The solid line shows the expected energy as a function of β; the area under this line is the value of F. The dashed lines show the upward path of a tempered transition, the dotted lines the downward path. The greater the area between these, the lower the acceptance probability.

The middle graph in Figure 1 shows a situation where the base transitions, $\hat{T}_i$ and $\check{T}_i$, do not produce immediate convergence. The typical result is that the energy is lower than it should be during the first half of the trajectory, during which the temperature is rising, and then exhibits a hysteresis effect in the second half of the trajectory, failing to fall immediately to the levels seen on the way up. As a result, $\check{F} - \hat{F}$ will tend to be large, reducing the acceptance probability. Because of this, the tempered transition method will work well only if the base transitions used are effective in establishing a good approximation to equilibrium. (If convergence is very slow, the acceptance probability might actually be quite high, simply because there was little change in the state, but if this is so, the tempered transitions will be ineffective in moving between modes anyway.)

Finally, the bottom graph in Figure 1 shows how tempered transitions lead to proper sampling when there are isolated modes. In the situation depicted, the base transitions for the low temperature distributions do not converge in a reasonable time to the true equilibrium distribution, though they may be adequate for sampling within a single mode. The curve traced during the first half of the tempered transition reflects the properties of the mode containing the initial state (up to the temperature at which movement between modes becomes possible). The curve traced during the second half of the tempered transition reflects the properties of the mode containing the final proposed state. The difference in the area under these two curves, which determines the probability of accepting the final state, relates to the relative probability mass in these two modes, with the result that the tempered transition method spends the proper amount of time in each mode even if the distribution at the high temperatures where movement between modes is possible is somewhat misleading. For the example shown in the figure, the transition will likely be rejected, even though the energy of the proposed state is lower than that of the current state, because the fall to this low energy in the second half of the tempered transition is delayed to quite low temperatures, an indication that the volume of the mode found is small. (Alternatively, this delayed fall could be due to slow convergence of the $\check{T}_i$, even within a mode.)

4 Performance when sampling from a multivariate Gaussian

To gain insight into the relative efficiency of simulated tempering and tempered transitions, we can look at their performance when sampling from a multivariate Gaussian distribution. Of course, we would not normally use these methods to sample from a Gaussian distribution, nor, for that matter, to sample from any unimodal distribution (except perhaps to confirm that it is indeed unimodal). Isolated modes of more interesting distributions will often resemble multivariate Gaussians over a certain range of temperatures, however, so we would not expect any method to perform well on realistic problems if it does not perform well on Gaussian distributions.

For simplicity of exposition, I will look at sampling from a distribution for $x = \langle x_1, \ldots, x_N \rangle$ in which the x_j are independent and each have a Gaussian distribution with mean zero and variance one. The results hold for any multivariate Gaussian, however, as long as suitable base transitions are provided that efficiently sample at the various temperatures. I will focus on high dimensional problems, where N is large.

The desired Gaussian distribution for x can be expressed as the canonical distribution at a temperature of one using the following energy function:

$$E(x) \;\; = \;\; \frac{1}{2} \sum_{j=1}^{N} x_i^2 \tag{19}$$

At other temperatures, the canonical distribution is also Gaussian, but with the x_j having a variance equal to the temperature, $T = 1/\beta$. The normalizing constants for these distributions (the "partition function") are given by

$$Z(\beta) = (2\pi/\beta)^{N/2} \tag{20}$$

Note that for large N, the canonical distribution for $E(x)$ at inverse temperature β_i will be approximately Gaussian, with mean $N/2\beta_i$ and variance $N/2\beta_i^2$.

Let us assume that we wish to apply simulated tempering or the tempered transition method to this problem using a highest temperature of $T_* = 1/\beta_*$, though there is of course no actual benefit in sampling at a high temperature for this problem. To compare the two methods, we need to determine the optimal number and spacing of the β_i that interpolate between $\beta_0 = 1$ and β_*. From this, we can find the efficiency with which each method moves between β_0 and β_*.

4.1 Simulated tempering for a multivariate Gaussian

For simulated tempering to work well, adjacent temperatures, given by β_i and β_{i+1}, must be close enough that movement between them occurs reasonably often — i.e. we must have a fairly high rate of acceptance both for proposals to change the current temperature from i to $i+1$ (while leaving the rest of the state, x, unchanged), and for proposals to change the temperature from $i+1$ to i. In evaluating these acceptance rates for the problem of sampling from a multivariate Gaussian, let us assume that we are using the weights $w_i = 1/Z(\beta_i) = (\beta_i/2\pi)^{N/2}$, for which the marginal distribution of i is uniform. Detailed balance then implies that the probability of accepting a proposed move from $i+1$ to i must be the same as that of accepting a move from i to $i+1$, so we can look at just the latter. From (3) and (20), we get the following expression for the probability of accepting a proposal to move from temperature i to temperature $i+1$ when the remainder of the state is x:

$$\min\Big[1,\ (\beta_{i+1}/\beta_i)^{N/2} \exp((\beta_i - \beta_{i+1})E(x))\Big] \;\; = \;\; \min\,[1, \exp(-D)] \tag{21}$$

where

$$D \;\; = \;\; (N/2)\, \log(\beta_i/\beta_{i+1}) \;\; - \;\; (\beta_i - \beta_{i+1})\, E(x) \tag{22}$$

If we assume that the transitions used to update x at each β_i immediately establish the corresponding canonical distribution, the distribution of D will have mean

$$\frac{N}{2}\left[\log(\beta_i/\beta_{i+1}) \;-\; \frac{\beta_i-\beta_{i+1}}{\beta_i}\right] \;=\; \frac{N}{2}\left[\log(\gamma)-1+\frac{1}{\gamma}\right] \;\approx\; N\delta^2/4 \tag{23}$$

and variance

$$\frac{N}{2}\left[\frac{(\beta_i-\beta_{i+1})^2}{\beta_i^2}\right] \;=\; \frac{N}{2}\left(1-\frac{1}{\gamma}\right)^2 \;\approx\; N\delta^2/2 \tag{24}$$

where $\gamma = 1+\delta = \beta_i/\beta_{i+1}$. The approximations hold when δ is small, as will be the case for β_i that work well when N is large.

Since the acceptance probability depends only on D, whose distribution depends only on β_i/β_{i+1}, the best way of spacing the β_i will be geometrical. We should choose the ratio of adjacent temperatures in the series to be as large as possible, while still retaining a reasonable acceptance rate. (I will not attempt here to find the exactly optimal acceptance rate; all that I need is for the optimal acceptance rate to reach a non-zero limit as N increases, which seems clear intuitively.)

There are two ways in which a reasonable acceptance rate could be ensured. One is for the expected value of D to be kept small in absolute terms. From (23), achieving this would require that we set $\delta \approx N^{-1/2}$. The other way is for the expected value of D to be comparable to its standard deviation, in which case D will often be negative, again ensuring a reasonable acceptance rate. From (23) and (24), this would also lead to the requirement that $\delta \approx N^{-1/2}$. This is therefore the scaling that we must adopt.

We conclude that for simulated tempering based on canonical distributions to work well when sampling from a Gaussian distribution of high dimensionality, N, it must use a series of distributions with inverse temperatures $1{=}\beta_0,\ \beta_1,\ \ldots,\ \beta_{n_{st}}{=}\beta_*$ for which $\beta_i/\beta_{i+1}-1 \approx N^{-1/2}$. The number of distributions used in addition to the one in which we are interested will therefore be $n_{st} \approx N^{1/2}\log\beta_*$. Since simulated tempering moves between these distributions via a random walk, the number of steps (each including a transition that updates x) that are needed to move from the distribution of interest (at $\beta_0{=}1$), to the highest temperature distribution (at β_*), and back to the distribution of interest will be around $2n_{st}^2 \approx 2N\log^2\beta_*$.

4.2 Tempered transitions for a multivariate Gaussian

The probability of accepting a tempered transition depends on the difference $\check{F}-\hat{F}$ (equation (13)), which from equations (14) and (15) is

$$\sum_{i=0}^{n-1}\,(\beta_i-\beta_{i+1})\,E(\check{x}_i) \;-\; \sum_{i=0}^{n-1}\,(\beta_i-\beta_{i+1})\,E(\hat{x}_i) \tag{25}$$

If we assume that the tempered transitions used to sample from a multivariate Gaussian distribution use base transitions, $\hat{T}_i$ and $\check{T}_i$, that establish their equilibrium distributions immediately, then for large N, the distribution of this difference will be Gaussian with mean

$$\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})\,\frac{N}{2\beta_{i+1}} \;-\; \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})\,\frac{N}{2\beta_i} \;=\; \frac{N}{2}\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})\left[\frac{1}{\beta_{i+1}} - \frac{1}{\beta_i}\right] \quad (26)$$

and variance

$$\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})^2\,\frac{N}{2\beta_{i+1}^2} \;+\; \sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})^2\,\frac{N}{2\beta_i^2} \;=\; \frac{N}{2}\sum_{i=0}^{n-1}(\beta_i - \beta_{i+1})^2\left[\frac{1}{\beta_{i+1}^2} + \frac{1}{\beta_i^2}\right] \quad (27)$$

The fact that each term in the above sums depends only on the ratio β_i/β_{i+1} indicates that here also we are best off using a geometric spacing for the inverse temperatures, with some constant $\gamma = 1 + \delta = \beta_i/\beta_{i+1}$. For small δ, the number of additional distributions used will then be $n_{tt} = \delta^{-1}\log\beta_*$. We can now rewrite the mean of $\check{F} - \hat{F}$ from (26) as follows, since all the terms in the sum are the same:

$$\frac{Nn_{tt}}{2}\,(\gamma - 1)\left(1 - \frac{1}{\gamma}\right) \;\approx\; N\log(\beta_*)\delta/2 \quad (28)$$

Similarly, we can write the variance from (27) as:

$$\frac{Nn_{tt}}{2}\,(\gamma - 1)^2\left(1 + \frac{1}{\gamma^2}\right) \;\approx\; N\log(\beta_*)\delta \quad (29)$$

As before, to maintain a reasonable acceptance rate, we must either keep the mean for $\check{F} - \hat{F}$ small in absolute terms, or we must keep the mean for $\check{F} - \hat{F}$ comparable to its standard deviation. Either requirement leads to $\delta \approx 1/(N\log\beta_*)$, and this must therefore be the way we scale δ. From this, we find that the number of additional distributions needed, n_{tt}, will be approximately $N\log^2\beta_*$.

The number of steps needed to move from the lowest temperature (at $\beta_0 = 1$), to the highest temperature (at β_*), and back to the lowest temperature using a tempered transition will therefore be about $2n_{tt} \approx 2N\log^2\beta_*$. This is essentially the same as was found for simulated tempering, though there might well be some constant factor difference that is not revealed by this analysis. The advantage that tempered transitions have in not doing a random walk is cancelled by the disadvantage that maintaining a reasonable acceptance rate requires that a larger number of distributions be used to interpolate between the lowest temperature and the highest temperature.

5 Performance on complex problems

The above analysis shows that the efficiency of the tempered transition method is similar to that of simulated tempering when sampling from a multivariate Gaussian; we may expect that performance will also be similar for other simple unimodal distributions. These are not the problems we are interested in, however. In this section, I will explore the performance of the two methods on more complex problems. I will first look at how the methods perform on a test problem that is simple enough to understand, but complex enough to perhaps be typical of more realistic problems. I will then discuss the more general significance of the possible "deceptive" aspects of high-temperature distributions, such as are seen in the test problem.

5.1 Performance on a test problem

To test the performance of simulated tempering and tempered transitions, I will use the problem of sampling from a two-dimensional distribution for $x = (x_1, x_2)$ that is an equally-weighted mixture of 4292 Gaussians. In all the component Gaussians, x_1 and x_2 are independent, and both have standard deviation 0.001. The means of the Gaussians place them in four groups, one in each quadrant of the plane. In the upper right quadrant are 121 Gaussians whose means are arranged in an 11×11 square array centred at $(+15, +15)$, with the spacing between means being 0.0025. In the upper left quadrant are 121 Gaussians whose means form an 11×11 square array centred at $(-15, +15)$, with spacing of 0.15. In the lower left quadrant are 2025 Gaussians whose means are arranged in a 45×45 square array centred at $(-15, -15)$, with the spacing between means being 0.0025. Finally, in the lower right quadrant are 2025 Gaussians in a 45×45 square array centred at $(+15, -15)$, with the spacing between means being 0.15.

It is, of course, quite easy to sample from this distribution given knowledge of how it is defined. The challenge is to efficiently sample from it knowing only that it is the canonical distribution at a temperature of one produced by the following "black box" energy function:

$$
\begin{aligned}
E(x) \;=\; -\log \Bigg[& \sum_{i=-5}^{+5} \sum_{j=-5}^{+5} \exp\left(-\frac{|x-\mu_{1,i,j}|^2}{2\sigma^2}\right) \;+\; \sum_{i=-5}^{+5} \sum_{j=-5}^{+5} \exp\left(-\frac{|x-\mu_{2,i,j}|^2}{2\sigma^2}\right) \\
& + \sum_{i=-22}^{+22} \sum_{j=-22}^{+22} \exp\left(-\frac{|x-\mu_{3,i,j}|^2}{2\sigma^2}\right) \;+\; \sum_{i=-22}^{+22} \sum_{j=-22}^{+22} \exp\left(-\frac{|x-\mu_{4,i,j}|^2}{2\sigma^2}\right) \Bigg]
\end{aligned}
\tag{30}
$$

where $\sigma = 0.001$, and

$$
\begin{aligned}
\mu_{1,i,j} &= (0.0025\,i + 15,\ 0.0025\,j + 15) \\
\mu_{2,i,j} &= (0.1500\,i - 15,\ 0.1500\,j + 15) \\
\mu_{3,i,j} &= (0.0025\,i - 15,\ 0.0025\,j - 15) \\
\mu_{4,i,j} &= (0.1500\,i + 15,\ 0.1500\,j - 15)
\end{aligned}
\tag{31}
$$

As the temperature is increased above one, the modes in the corresponding canonical distributions based on this energy function spread out. They merge into a single mode at a temperature of about $T = 2^{28}$, and this will therefore be the highest temperature at which sampling is done. Figure 2 shows samples from a series of distributions at different temperatures (obtained from a simulated tempering run in which good sampling was ensured by including a special transition for $T = 1$ that samples directly from the true distribution). Note that these are *not* the same distributions as would be produced by varying σ. In particular, the total probability of the states in each of the four quadrants changes as the temperature changes, an effect produced by the different spacings of modes

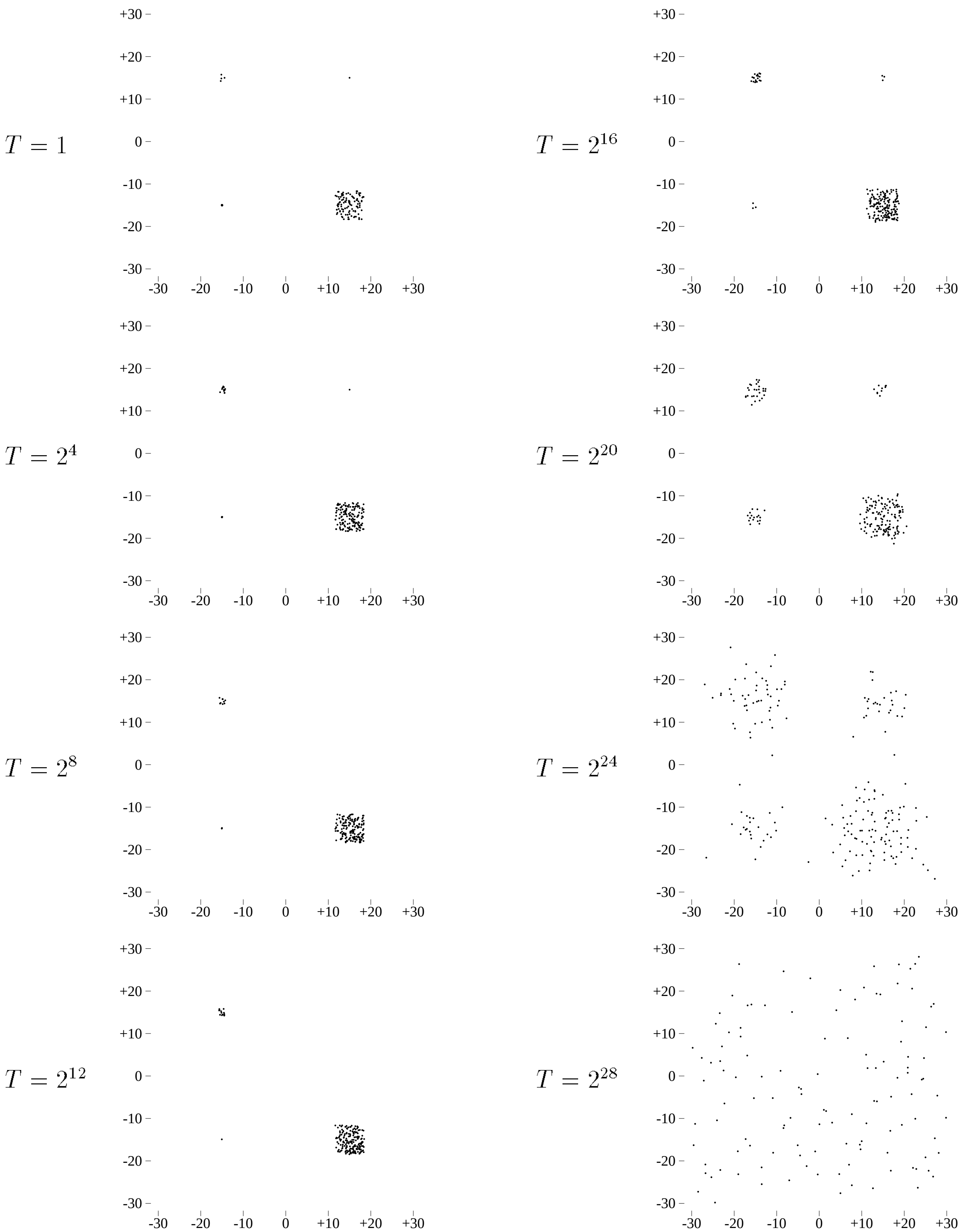

Figure 2: Some canonical distributions for the test problem. Each plot shows 200 points drawn from the distribution at the indicated temperature (except that 78 points are off the plot for $T = 2^{28}$). For $T = 1$, there are actually about the same number of points in the lower-left and lower-right, the ones in the lower-left are just close together. For $T = 2^8$ and $T = 2^{12}$, however, there are very few points in the lower-left.

within the arrays making up the groups in each quadrant.

To sample from one of these canonical distributions, at a temperature of T_i, I used base transitions consisting of ten Metropolis updates, for which the proposal distribution was a bivariate Gaussian with the current value of x as the mean and a covariance matrix of $\sigma^2 T_i \mathbf{I}$. The acceptance rate for such proposed changes to x was 70%, averaged over all temperatures used, and was similar to this for each individual temperature. This therefore appears to be a reasonable method of sampling from these distributions; whether it is the best method is not important for testing tempering methods. Note that at low temperatures these Metropolis updates are very unlikely to jump from the mode where the current state is located to a distant mode. This is typical of the problems on which the tempering methods are meant to be used.

For the simulated tempering runs, the Metropolis updates for x were alternated with Metropolis updates for the temperature index, i, as described in Section 2. These runs sampled from eight canonical distributions at temperatures of $T_i = 2^{4i}$, for $i = 0, \ldots, 7$; these are the distributions shown in Figure 2. As we saw, this geometric spacing for the T_i is optimal for sampling from a multivariate Gaussian. It may not be optimal for this problem, but I did not attempt to find a better sequence of temperatures. Eight was chosen as an appropriate number of geometrically-spaced temperatures by trial and error. The weights (w_i) for these distributions were also selected by trial and error, based on the results of preliminary runs. The weights eventually chosen were as follows:

$$\begin{array}{rclcrcl} \log(w_0) &=& 0 & \quad & \log(w_4) &=& -7.6 \\ \log(w_1) &=& -2.2 & & \log(w_5) &=& -8.3 \\ \log(w_2) &=& -4.8 & & \log(w_6) &=& -10.0 \\ \log(w_3) &=& -7.0 & & \log(w_7) &=& -12.2 \end{array} \tag{32}$$

The amount of time spent at each of the eight temperatures in the simulated tempering runs that used these weights was uniform to within a factor of about 1.5. The acceptance rate for proposed changes to the temperature index, i, was 35%, indicating that the spacing between temperatures was neither excessively large nor excessively small.

For the runs using tempered transitions, I used 200 temperatures, geometrically spaced, ranging from $T_0 = 1$ to $T_{199} = 2^{28}$. Here again, this geometric spacing of temperatures may not be optimal. The number of temperatures used was chosen by trial and error so as to be as small as possible while keeping the acceptance rate reasonably high. Using 200 temperatures, the acceptance rate was 30%. Before each tempered transition, I did 20 simple Metropolis updates of x (at a temperature of one), mostly to avoid having points from successive iterations coincide exactly when a tempered transition is rejected (which would make scatterplots hard to interpret).

The total number of Metropolis updates for x that are performed in one such tempered transition together with the 20 Metropolis updates preceding it is $20+2\times 199\times 10 = 4000$, which is the same as the number of updates for x in 400 simulated tempering iterations. Since these updates for x are the major contributor to the computation time, both in this

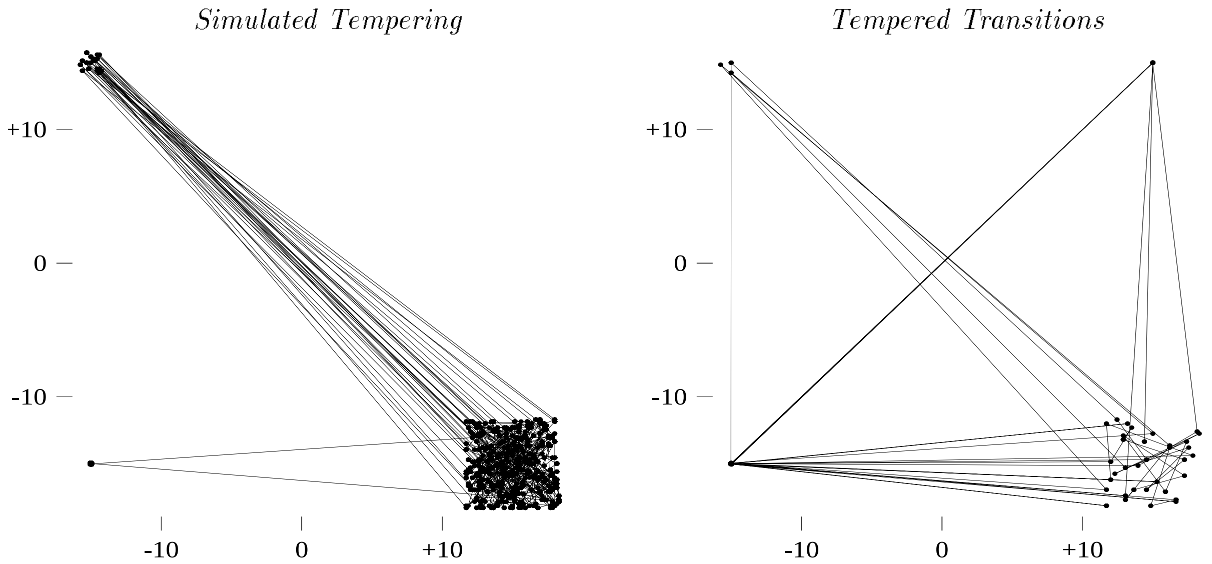

Figure 3: Results of simulated tempering and the tempered transition method applied to the test problem. The connected dots show the successive states that were obtained from the distribution at $T = 1$.

test problem and in typical real problems, a fair comparison of the two methods should allow the simulated tempering runs to continue for 400 times as many iterations as the runs using tempered transitions. I did runs of 200 tempered transitions and of 80 000 simulated tempering iterations.

The results obtained using simulated tempering and tempered transitions are shown in Figure 3. The points plotted on the left are the 9187 states out of the 80 000 in the simulated tempering run for which i was zero, and hence T_i was one. The points plotted on the right are the states following each of the 200 tempered transitions (for some of which the candidate state was rejected).

The two methods clearly perform quite differently on this problem. The simulated tempering run moved easily between and within the upper left and lower right quadrants. It took only a single excursion into the lower left quadrant, however, though this excursion accounts for a significant fraction of the entire sample of states at $T = 1$: 4356 states out of 9187 (this is not visible from the figure, as these states are not resolved at the scale of the plot). During this excursion, the simulated tempering run sampled well from the conditional distribution for states within the lower left quadrant, but this single excursion provides little basis for estimating the total probability of states in the lower left. The simulated tempering run failed to produce any states at all from the upper right quadrant, whose states have a total probability of 2.8%.

In contrast, the tempered transitions moved many times between all four quadrants, providing a good basis for estimating the probability of being in each quadrant, as well as the expectations of other functions of x. Sampling was also adequate within each quadrant, though not as good as was seen with simulated tempering. Such short-range sampling might be improved by using additional tempered transitions with a lower maximum temperature (and hence a lower computational cost).

The failure of simulated tempering to perform adequately on this problem can be traced to the deceptive nature of the canonical distributions at $T=2^8$ and $T=2^{12}$, under which the states in the lower left and upper right have much smaller probability than they have under the distribution of interest, at $T=1$. This barrier blocks movement to or from low temperature states when x is in these quadrants. Putting it another way, there are three isolated regions of high probability in the full simulated tempering state space: one around $i=0$ and $x=(-15,-15)$, another around $i=0$ and $x=(+15,+15)$, and a third that covers the other two quadrants at low temperatures plus all four quadrants at high temperatures. One might hope to avoid this isolation by allowing i to change by more than ± 1 in a single iteration, perhaps using Gibbs sampling rather than a Metropolis update. This will not work well however — since the distribution for points within a quadrant at the high temperatures above the barrier is much broader than the distribution at low temperatures below the barrier, it is unlikely that a point drawn from the high temperature distribution will be in an area of high probability under the low temperature distribution, which is what would be needed for a jump to this low temperature to occur.

5.2 Effects of deceptive distributions on tempering schemes

The preceding example shows that simulated tempering can perform poorly when the additional distributions used to assist sampling are "deceptive". This should not be a surprise — there is no magical way by which efficient sampling can be guaranteed in all difficult problems. Tempering methods work only to the extent that the additional distributions they use are good guides to the distribution of interest. The less obvious conclusion from this example is that simulated tempering and tempered transitions can differ in how they are affected by the deceptive aspects of the distributions used. In the test problem, tempered transitions performed better, but we will see below that the advantage can also go the other way.

I will make a preliminary attempt here to clarify how the distributions used in tempering schemes can be deceptive. From any such distribution, p_i, we can derive an "associated final distribution", which is the distribution obtained by drawing a state, x, at random from p_i and then performing an annealing run, in which we successively update x using the base transitions associated with p_{i-1}, p_{i-2}, $\ldots$, p_0. The associated final distribution will be close to p_0 if these transitions are capable of moving between modes at low temperatures, but presumably this is not the case if we are using a tempering scheme.

Let us call a distribution "deceptive" if its associated final distribution differs substantially from p_0. Distributions can be deceptive in at least two ways. If some states with non-negligible probability under p_0 have much lower probability in this associated final distribution, I will say that the distribution is "unfaithful". If the associated final distribution gives a significant probability to states that have much lower probability under p_0, I will say that the distribution is "undiscriminating". We would like to relate the performance of tempering schemes on a problem to the ways in which the p_i used for the problem are deceptive.

Simulated tempering will work well as long as none of the p_i used are unfaithful, since movement will then be possible from the high temperature distribution (where sampling is assumed to be efficient) to any high-probability region of p_0. If some of the p_i are undiscriminating, some time will be wasted exploring paths that lead nowhere, but this effect should not be disastrous. If, on the contrary, some of the p_i are unfaithful, as in the test problem, simulated tempering can be very inefficient. It is possible, however, that there are cases where some of the p_i are unfaithful, as defined here, but simulated tempering still works well, because of the possibility that paths from an unfaithful p_i to all parts of p_0 might exist in which the temperature does not necessarily decrease monotonically.

The tempered transition method should work well as long as the p_i for the temperatures where modes divide are not too deceptive. In the test problem, for example, we see that the distribution changes from having a single mode at $T=2^{28}$ to having four modes at $T=2^{24}$. If the distribution at this point were unfaithful (which it is not), there would be a region of significant probability under p_0 that is only rarely reached in the second half of a tempered transition, and sampling would not be efficient. (Note, however, that when modes split at several temperatures, all that is actually required is that the conditional probabilities for each of the new modes given that one is in the region of the old unsplit mode be faithful; whether the unconditional probabilities are faithful is irrelevant.) If there is only a single temperature at which modes divide, an undiscriminating distribution will lead to only a modest wastage of time (as tempered transitions are occasionally directed to dead ends). If, however, there are many temperatures at which modes split, the waste can grow exponentially with the number of splits, if each suffers from a lack of discrimination. In the test problem, the modes split twice. For the first split (at $T=2^{24}$), there is a wastage of about a factor of two; for the second split (at low temperature), the distributions are not undiscriminating.

We see, therefore, that tempered transitions may work better than simulated tempering when, as in the test problem, the distributions used are unfaithful at some temperatures, but not at the temperatures where modes split. Simulated tempering may work better than tempered transitions if modes split successively at many temperatures, with the distributions at these temperatures being faithful, but to some degree undiscriminating.

6 Discussion

I have shown that tempered transitions can be used to explore multimodal distributions with an efficiency that on simple problems is comparable to that of simulated tempering. When applied to complex distributions, the two methods can differ significantly in performance; it is not clear which will perform best on typical problems. The major advantage of the tempered transition method is that it does not require approximate values of the normalizing constants for the p_i. Producing such estimates can be a tedious process. (However, simulated tempering does in the end produce good estimates of these normalizing constants, which may sometimes be of interest.)

It is disappointing, however, that while the tempered transition method avoids doing a random walk between distributions, this advantage is cancelled (up to a possible constant factor) by the greater number of distributions needed in order to produce a reasonable acceptance probability. One should also note that simulated tempering allows data associated with the p_i other than p_0 to be used to calculate expectations with respect to the other p_i, which may sometimes be of interest, and to a limited extent with respect to p_0 (using an importance sampling estimator). Simulated tempering also allows use of "regenerative" simulation methods (Geyer and Thompson 1994), though it is unclear whether these provide any real advantage.

The trade-offs evident in the above comparison prompt one to wonder whether it is possible to devise a tempered sampling scheme that combines all or many of the advantages of the various present schemes. We can characterize tempering schemes as follows:

1) How many interpolating distributions are required in order to produce a reasonable acceptance rate? In particular, when sampling from a multivariate Gaussian, is the number, n, of distributions required proportional to $N^{1/2}$, or to N?

2) Does information from states sampled at high temperatures move to lower temperatures via a systematic process, taking around n steps, or via a random walk, taking around n^2 steps?

3) Does one step of the movement described above require only one (or a few) basic transitions, or is a larger number (e.g. n) required?

4) Does the procedure require that preliminary estimates of normalization constants for the p_i be found?

5) Does the procedure produce good estimates of the normalizing constants for the p_i as a by-product of the sampling?

6) Does the procedure allow expectations to be found with respect to all the p_i, or only with respect to p_0?

7) Does the procedure require storing only a single state at any time, or is storage required for n states, one for each p_i?

The three tempering schemes so far devised can be scored on these questions as follows:

	1	2	3	4	5	6	7
Simulated Tempering	$\surd$		$\surd$		$\surd$	$\surd$	$\surd$
Metropolis Coupled Markov Chains	$\surd$			$\surd$	$\surd$	$\surd$	
Tempered Transitions		$\surd$	$\surd$	$\surd$			$\surd$

Here, a $\surd$ in column c means that the answer to question c above is that which is obviously to be preferred. The answers to the first three questions determine the speed of sampling, which is perhaps the most important characteristic for most problems. In this respect, simulated tempering and tempered transitions have the advantage over Metropolis coupled Markov chains (Geyer 1991). However, the latter method is the only one that produces good estimates of normalization constants without requiring preliminary estimates to start.

By using $\hat{T}_i$ and $\check{T}_i$ that "over-relax" the energy, it is possible to create a version of the tempered transition method that works well when the number of interpolating distributions is similar to that needed by simulated tempering (question (1)). Unfortunately, the computation per step rises by a corresponding amount (question (3)). I have also devised a version of simulated tempering that avoids doing a random walk (question (2)), but again the increased computation per step (question (3)) cancels the benefit. There is no apparent reason why attempts of this nature must necessarily fail, however. I also hope that a method combining ideas from Metropolis coupled Markov chains and tempered transitions will score well on questions (1), (2), (4), (5), and (6), and will therefore be as good as or better than existing methods with respect to all characteristics except space utilization.

In another direction, I will briefly note here that the general concept of tempered transitions can be applied to the hybrid Monte Carlo algorithm of Duane, Kennedy, Pendleton, and Roweth (1987), in the form of "tempered trajectories". The hybrid Monte Carlo algorithm is an elaborate form of the Metropolis algorithm, operating in an extended "phase space" that includes "momentum" variables, in which candidate states are proposed by computing a dynamical trajectory, usually based on the Hamiltonian dynamics that corresponds to the distribution being sampled.

Validity of the hybrid Monte Carlo algorithm does not depend on using Hamiltonian dynamics, however — all that is required is that the trajectories be reversible and preserve phase space volume. We can therefore modify the algorithm as follows: During the first half of each trajectory, we multiply all the momentum variables by some constant slightly greater than one after performing each discretized step of the dynamics; during the second half of the trajectory, we divide all the momentum variables by the same constant before performing each step of the dynamics. The multiplications of the momentum expand phase space volume in the first half, but this expansion is exactly cancelled by the contraction in the second half, leaving phase space volume unchanged for the trajectory as a whole.

The effect of this procedure is to increase the magnitude of the momentum in the middle part of the trajectory. This effectively imposes a higher "temperature" on the system, which leads it to explore a wider region of state space, and perhaps move from one mode to another. The changes in temperature are not pre-determined with this technique, but rather depend on the "specific heat" of the system, in a manner that is probably beneficial.

In preliminary experiments, I have found that the number of intermediate distributions that are effectively required when using tempered trajectories is less than are needed with tempered transitions, but this may simply reflect a constant factor advantage, since rough theoretical arguments indicate that the scaling behaviour of the two methods should be similar.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the Information Technology Research Centre.

Bibliography

Berg, B. A. and Celik, T. (1992) "New approach to spin-glass simulations", *Physical Review Letters*, vol. 69, pp. 2292-2295.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987) "Hybrid Monte Carlo", *Physics Letters B*, vol. 195, pp. 216-222.

Geyer, C. J. (1991) "Markov chain Monte Carlo maximum likelihood", in E. M. Keramidas (editor), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156-163, Interface Foundation.

Geyer, C. J. and Thompson, E. A. (1994) "Annealing Markov chain Monte Carlo with applications to ancestral inference", Technical Report No. 589 (revised February 7, 1994), School of Statistics, University of Minnesota.

Kennedy, A. D. (1990) "The theory of hybrid stochastic algorithms", in P. H. Damgaard, *et al.* (editors) *Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, New York: Plenum Press.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983) "Optimization by simulated annealing", *Science*, vol. 220, pp. 671-680.

Marinari, E. and Parisi, G. (1992) "Simulated tempering: A new Monte Carlo Scheme", *Europhysics Letters*, vol. 19, pp. 451-458.

Neal, R. M. (1993) "Probabilistic inference using Markov Chain Monte Carlo methods", Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Smith, A. F. M. and Roberts, G. O. (1993) "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods" (with discussion), *Journal of the Royal Statistical Society B*, vol. 55, pp. 3-23 (discussion, pp. 53-102).