Technical Report No. 9910, Department of Statistics, University of Toronto

# Circularly-Coupled Markov Chain Sampling

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
`http://www.cs.utoronto.ca/~radford/`
`radford@stat.utoronto.ca`

22 November 1999

**Abstract.** I show how to run an $N$-time-step Markov chain simulation in a circular fashion, so that the state at time 0 follows the state at time $N-1$ in the same way as states at times $t$ follow those at times $t-1$ for $0<t<N$. This wrap-around of the chain is achieved using a coupling procedure, and produces states that all have close to the equilibrium distribution of the Markov chain, under the assumption that coupled chains are likely to coalesce in less than $N/2$ iterations. This procedure therefore automatically eliminates the initial portion of the chain that would otherwise need to be discarded to get good estimates of equilibrium averages. The assumption of rapid coalescence can be tested using auxiliary chains started at times spaced between 0 and $N$. When multiple processors are available, such auxiliary chains can be simulated in parallel, and pieced together to give the circularly-coupled chain, in less time than a sequential simulation would have taken, provided that coalescence is indeed rapid. The practical utility of these procedures is dependent on the development of good coupling schemes, but in contrast to exact sampling techniques such as coupling from the past, there is no need to also devise a way of keeping track of large sets of states. On the other hand, although the assumptions behind circular coupling can be tested empirically, the results will not provide an absolute guarantee that the points obtained are from the equilibrium distribution.

## 1 Introduction

Sampling from a complex distribution by simulating a Markov chain having this distribution as its equilibrium distribution is an important technique in statistical mechanics (eg, Frenkel and Smit 1996), in Bayesian statistics (eg, Gilks, *et al* 1996), and in other computational problems (eg, Sinclair 1993). States from the equilibrium distribution of the chain are then used to estimate quantities of interest, such as the average energy of a physical system, or a Bayesian predictive distribution.

Ideally, such simulations would be conducted with theoretical knowledge of the time needed for the chain to reach its equilibrium distribution to within a given tolerance. Although some progress has been made at producing quantitative bounds on convergence times of Markov chains (eg, Rosenthal 1995a), such theoretical guarantees are presently unavailable for most problems of practical interest.

Instead, practitioners usually assess the convergence of the Markov chain sampler empirically, by formal or informal methods. These methods attempt to determine whether the chain has reached an adequate approximation to its equilibrium distribution within the number of iterations simulated. If it appears to have done so, some initial portion of the chain (the "burn-in" period) is generally discarded in order to avoid biasing the results by inclusion of states that reflect the state in which the chain was started rather than its equilibrium distribution. The bewildering variety of methods for diagnosing convergence and discarding an appropriate burn-in period have been reviewed by Cowles and Carlin (1996), Brooks and Roberts (1998), and Mengersen, *et al* (1999).

One convergence diagnostic, due to Johnson (1996), looks at multiple "coupled" chains that are started from different initial states, but that subsequently undergo transitions determined by the same random numbers. Rosenthal (1995b) reviews the application of coupling to Markov chains. Briefly, coupling chains introduces dependencies between them, and may lead them to "coalesce" to the same state after some number of iterations. The probability that a chain started from the initial state distribution has not coalesced with a chain started from the equilibrium distribution by time $T$ provides an upper bound on the total variation distance of the chain from equilibrium at time $T$. In Johnson's diagnostic, the time required for several chains whose initial states were drawn from a distribution that is meant to be "overdispersed" with respect to the equilibrium distribution is taken as an informal indication of how much time is required for approximate equilibrium to be reached.

One way of viewing the circular coupling method of this paper is as a refinement of Johnson's scheme, which addresses two problems that scheme suffers from. One problem noted by Johnson is that using the states immediately following the time when all chains coalesce introduces a bias in the results, favouring states where coalescence is more likely. The circular coupling scheme of this paper discards an initial portion of the chain without introducing bias (under certain assumptions), by using the last state of the chain to start a new chain at time zero, and using the states of this chain rather than of the original chain up to the point where it and the original chain coalesce. Another problem is that although Johnson's scheme considers several initial states, it uses only a single sequence of random numbers. It could be that this one sequence happens to produce atypically fast coalescence. Diagnostics in the circular coupling scheme are based on a variety of starting states at times spaced throughout the total simulation period, thereby effectively considering various initial random number sequences. This reduces (but does not eliminate) the possibility that the results will be misleadingly optimistic.

The coupling technique of this paper also provides a way of exploiting parallel computation in Markov chain simulation. As discussed by Rosenthal (1999), there are many possibilities for exploiting multiple processors for the overall task of estimating quantities using Markov chain Monte Carlo methods. However, the core operation of simulating a single realization of a Markov chain might appear to be inherently sequential, since it might seem that the state at time $t$ cannot be obtained until the state at time $t-1$ has been found. In this paper, however, I will show how coupling allows one to use multiple processors to simulate a single realization of a Markov chain in substantially less time that would be required by a single processor, provided that the Markov chain and the coupling method employed lead to rapid coalescence of chains.

The practical feasibility of circular coupling is dependent on finding an efficient coupling scheme — ie, a way of introducing dependencies between the transitions from chains currently in different states that promotes the rapid coalescence of these chains to the same state. General techniques for achieving this exist, one of which will be introduced here and used for some simple demonstrations. However, there may well be a cost to restricting the Markov chains used to those for which good coupling methods are available, and even when a suitable coupling scheme is available, it may not be optimal, and hence may lead to coalescence times that are greater than the actual time required for the Markov chain to reach approximate equilibrium.

This need for an efficient coupling scheme is less onerous than for the alternative of exact (a.k.a. "perfect") sampling methods, such as coupling from the past (Propp and Wilson 1996) and the interruptible scheme of Fill (1998). For these methods, the coupling scheme must not only promote coalescence, but also permit the efficient tracking of large sets of states, so that the coalescence of a huge (possibly infinite) set of chains started in all possible states can be determined. Circular coupling looks only at the coalescence of a moderate number of explicitly simulated chains, and is therefore much easier. The price of this is that circular coupling will not provide an absolute guarantee that the states obtained are from the exact equilibrium distribution, but only an assurance that they are from close to the equilibrium distribution, provided that certain conditions are met whose truth can be empirically tested, but not verified with certainty.

## 2 The basic circular coupling procedure

Suppose we wish to sample from a distribution $\pi$ for some state variable $x$ by using a Markov chain having $\pi$ as an invariant distribution. We hope that this Markov chain is ergodic, and hence has $\pi$ as its only invariant distribution, and that it converges to this equilibrium distribution rapidly.

Realizations of this chain with different initial states can be coupled by representing the transitions of the chain by a function $\phi(x_t, u_t)$, which takes as arguments the state at some time, $x_t$, and the random numbers generated at that time, $u_t$, and returns the state of the chain at the next time, $x_{t+1}$. The random numbers at each time are drawn independently, each from the same distribution, $U$. There will be many ways of expressing a given set of transition probabilities in this way, using different transition functions, $\phi$, and different random number distributions, $U$. These different ways of coupling the chains may lead to coalescence occurring more or less rapidly. However, with any coupling scheme of this nature, once two chains have coalesced to the same state at some time, they will remain in the same state at all subsequent times.

With this representation, an ordinary Markov chain simulation for $N$ time steps is conducted as follows:

**Standard Markov chain simulation:**

1) Randomly draw $x_0$ from the initial state distribution, $p_0$.

2) For $t = 1, \ldots, N$:

   Randomly draw $u_{t-1}$ from the distribution $U$, independently of previous draws.
   Let $x_t = \phi(x_{t-1}, u_{t-1})$.

A circularly-coupled Markov chain simulation for $N$ time steps begins the same way as a standard simulation, but after generating $x_0, \ldots, x_N$, a second set of states, $y_0, \ldots, y_N$, are generated by letting $y_0 = x_N$ and then redoing the simulation from this starting point, using the same random numbers, $u_0, \ldots, u_{N-1}$, as before. If the chain started from $y_0$ coalesces with the original chain before time $N$, there is no need to perform any further computations, since each $y_t$ from that time on will be the same as the corresponding $x_t$. Here is the basic procedure (without the diagnostics that will be added in Section 4), which is also illustrated in Figure 1:

**Basic circularly-coupled Markov chain simulation:**

1) Randomly draw $x_0$ from the initial state distribution, $p_0$.

2) For $t = 1, \ldots, N$:

   Randomly draw $u_{t-1}$ from the distribution $U$, independently of previous draws.
   Let $x_t = \phi(x_{t-1}, u_{t-1})$.

3) Let $y_0 = x_N$.

4) For $t = 1, \ldots, N$ while $y_{t-1} \neq x_{t-1}$:

   Let $y_t = \phi(y_{t-1}, u_{t-1})$.

5) Let the remaining $y_t$ be the same as the corresponding $x_t$.

In practice, pseudo-random numbers would generally be used, eliminating the need to save $u_0, \ldots, u_{N-1}$ — instead, the pseudo-random number generator can simply be re-initialized using the original seed. If the amount of memory needed to save a state is large, the values of $x_0, \ldots, x_{N-1}$ might not be saved when they are first generated, but instead be recomputed as $y_0, y_1, \ldots$ are generated. This approach might be preferable when coalescence of $y_0, y_1, \ldots$ with $x_0, x_1, \ldots$ is expected to be rapid. Of course, the values of functions of the state that are of interest will still need to be saved, so that estimates can in the end be computed for the expectations of these functions.

If the wrapped-around chain fails to coalesce with the original chain (ie, if $y_N \neq x_N$), the procedure may be seen as having failed. The project of sampling from $\pi$ might then

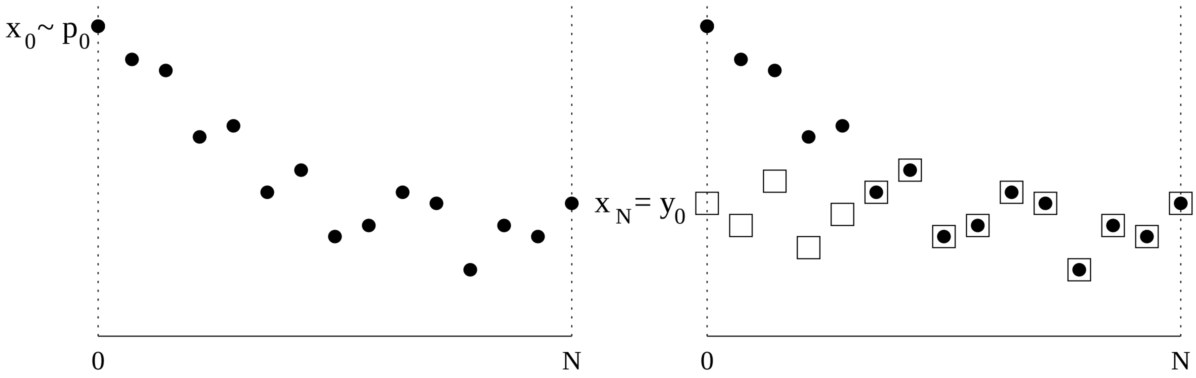


Figure 1: The basic circular coupling procedure. In this illustration, the real-valued state is plotted on the vertical axis, and simulation time is on the horizontal axis. The plot on the left shows the generation of the original chain, $x_0, \ldots, x_N$, starting with a state drawn from the distribution $p_0$. The plot on the right shows the subsequent generation of the wrapped-around chain, $y_0, \ldots, y_N$. In this case, the two chains coalesce at $t = 5$.

be abandoned, or the procedure might be redone with a substantially larger value for $N$. For purposes of theoretical analysis, however, I will assume that the $y_t$ for $t = 0, \ldots, N-1$ are always treated as a sample from the equilibrium distribution of the chain, and used to estimate expectations of functions with respect to this distribution, even if coalescence did not occur. (Note that since $y_N = y_0$ when coalescence does occur, these points should not both be included in the sample, as this point would then count double.) In the next section, I show that all these $y_t$ will indeed have approximately the equilibrium distribution, provided certain assumptions regarding speed of coalescence are satisfied. Of course, the $y_t$ will generally be dependent, and this will need to be accounted for when assessing the accuracy of the estimates obtained, as with standard Markov chain Monte Carlo procedures.

## 3 Proof of approximate correctness

The following theorem guarantees the approximate correctness of circular coupling, under certain assumptions:

**Theorem:** *Each point $y_t$, for $t = 0, \ldots, N$, that is generated by the basic circularly-coupled Markov chain simulation procedure with a given N (assumed here to be even) has a distribution that is within $2\epsilon + \delta$ of the equilibrium distribution, $\pi$, in total variation distance,[1] provided $\epsilon$ and $\delta$ are such that the following conditions hold regarding coupled chains (ie, chains that are simulated using the same random number sequence, $u_0, u_1, \ldots$):*

1) *If two chains are started from states drawn from the equilibrium distribution, $\pi$, independently of each other, and of the $u_t$, they will coalesce within $N/2$ iterations with probability at least $1 - \epsilon$.*

2) *If a chain is started from a state drawn from $\pi$, independently of the $u_t$, and another chain is started from a state drawn from the distribution $p_0$, independently of the initial*

[1] Here, the total variation distance between distributions $\mu$ and $\nu$ is defined to be $\sup_A |\mu(A) - \nu(A)|$, where the supremum is over all events $A$. Total variation distance is sometimes given an alternative definition that is twice this.

*state of the other chain and of the $u_t$, then the two chains will coalesce within $N$ iterations with probability at least $1-\delta$.*

The proof looks at another way of generating the sequence $y_0,\ldots,y_N$, along with $x_0,\ldots,x_N$, by means of the procedure given below, and illustrated in Figure 2.

**Theoretical circular coupling procedure:**

1) Randomly draw $x_0$ from the initial state distribution, $p_0$.

2) For $t=1,\ldots,N$:

   Randomly draw $u_{t-1}$ from the distribution $U$, independently of previous draws.
   Let $x_t = \phi(x_{t-1}, u_{t-1})$.

3) Randomly draw $v_0$ and $w_{N/2}$ from $\pi$, each independently of the other and of the $u_t$.

4) For $t=1,\ldots,N/2$: Let $v_t = \phi(v_{t-1}, u_{t-1})$.
   For $t=N/2+1,\ldots,N$: Let $w_t = \phi(w_{t-1}, u_{t-1})$.

5) Let $v^*_{N/2} = v_{N/2}$ and $w^*_0 = w_N$.

6) For $t=N/2+1,\ldots,N$: Let $v^*_t = \phi(v^*_{t-1}, u_{t-1})$.
   For $t=1,\ldots,N/2$: Let $w^*_t = \phi(w^*_{t-1}, u_{t-1})$.

7) Let $y_t = w^*_t$ for $t=0,\ldots,N/2-1$ and let $y_t = v^*_t$ for $t=N/2,\ldots,N$.

The generation of $x_0,\ldots,x_N$ in steps (1) and (2) above is the same as for the practical circular coupling procedure of the previous section, but the rest of the theoretical procedure could not be carried out in practice, since it requires sampling directly from $\pi$, which is presumably infeasible.

However, we can use this theoretical procedure to prove the approximate correctness of the practical circular coupling procedure. First, we will see that the $y_t$ produced by the theoretical procedure all have distribution $\pi$. Second, when the two conditions of the theorem hold, we will see that the distribution of $y_t$ obtained with the theoretical procedure is approximately the same as for the practical procedure.

That each of the $y_t$ obtained using the theoretical procedure has distribution $\pi$ follows easily as long as $\pi$ is an invariant distribution of the Markov chain defined by the function $\phi$ and the random number distribution $U$ — ie, as long as the distribution of $\phi(x,u)$ is $\pi$ when $x$ has distribution $\pi$ and $u$ has distribution $U$, independent of $x$. From this, it follows that all the $v_t$ and $v^*_t$ and all $w_t$ and $w^*_t$ have distribution $\pi$, since they are produced by transitions that leave $\pi$ invariant, starting from a state drawn from $\pi$. Since each $y_t$ is defined to be equal to either $w^*_t$ or $v^*_t$, the $y_t$ must all have distribution $\pi$ as well.

For the second part of the proof, we first note that when the two conditions of the theorem hold, the following events occur with the indicated probabilities:

a) $v_{N/2} = w^*_{N/2}$ with probability at least $1-\epsilon$.
b) $w_N = v^*_N$ with probability at least $1-\epsilon$.
c) $x_N = v^*_N$ with probability at least $1-\delta$.

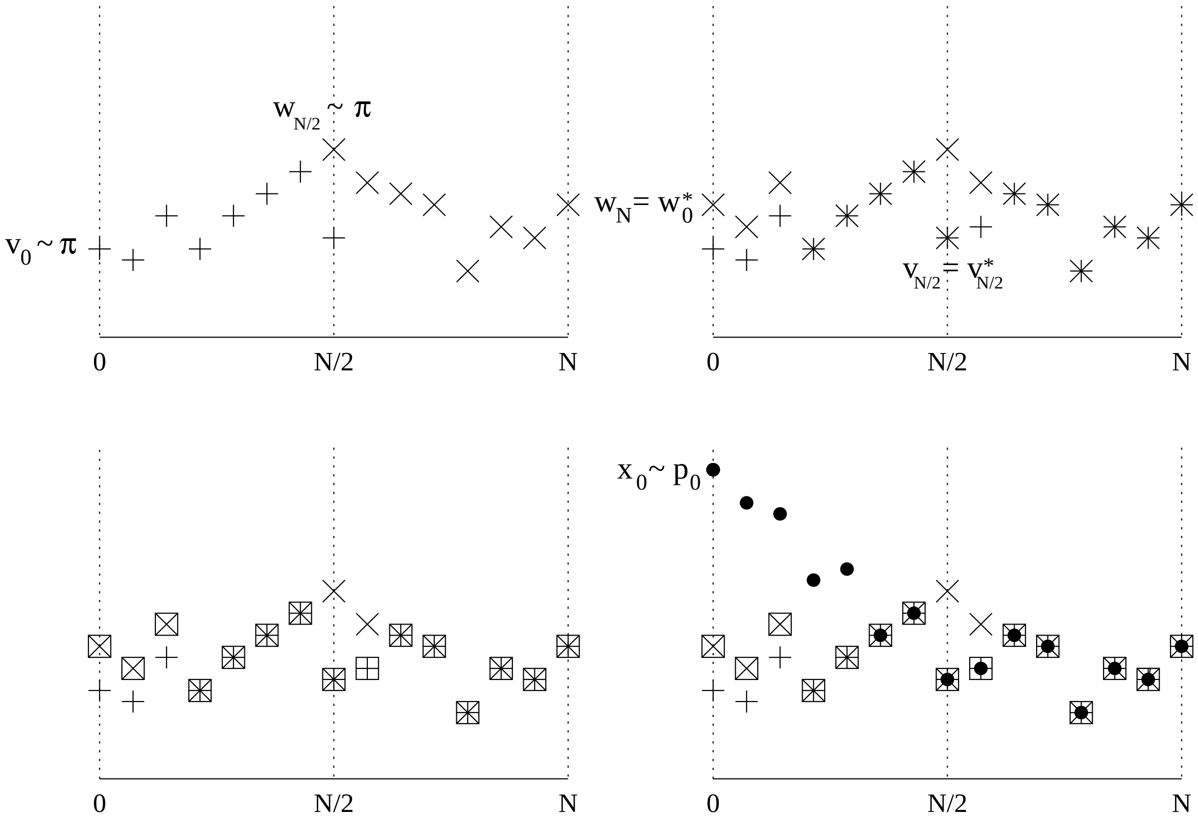


Figure 2: Proof of approximate correctness. The $v_t$ and $v_t^*$ are shown as $+$ signs, the $w_t$ and $w_t^*$ as $\times$ signs, the $y_t$ as squares, and the $x_t$ as dots. The top left shows generation of $v_0, \ldots, v_{N/2}$ and $w_{N/2}, \ldots, w_N$ starting from states drawn from $\pi$. The top right shows the continuation of these sequences, $v_{N/2} = v^*_{N/2}, \ldots, v^*_N$ and $w_N = w^*_0, \ldots, w^*_{N/2}$. In the bottom left, the sequence $y_0, \ldots, y_N$ used for estimation is identified. The bottom right shows a sequence $x_0, \ldots, x_N$ started from $p_0$ coalescing with the sequence $y_0, \ldots, y_N$, permitting this sequence to be found without the need to sample from $\pi$.

For event (a), this follows from Condition (1) because $v_{N/2}$ and $w^*_{N/2}$ are the result of $N/2$ iterations with the same $u_t$ (for $t = 0, \ldots, N/2-1$), starting from points $v_0$ and $w^*_0$ that are both drawn independently from $\pi$, independently of these $u_t$. That $w^*_0 = w_N$ is drawn from $\pi$ follows from it being produced from $w_{N/2}$, which is drawn from $\pi$, by applying transitions that leave $\pi$ invariant. Note that these transitions are defined in terms of $u_t$ for $t = N/2, \ldots, N-1$, which does not overlap the set of $u_t$ above. The bound on the probability of event (b) follows from Condition (1) similarly, and the bound on the probability of event (c) follows from Condition (2).

The event of (a), (b), and (c) all occurring has probability at least $1 - 2\epsilon - \delta$, because the probability that at least one of (a), (b), and (c) will not occur is at most $2\epsilon + \delta$ (this bound is valid even though the three events may not be independent). When this happens, $x_N = v^*_N = w_N = w^*_0 = y_0$. Furthermore, $y_t = \phi(y_{t-1}, u_{t-1})$ for $t = 1, \ldots, N$ — this is obvious for all $t$ except $N/2$, and for that $t$, it is a consequence of $v^*_{N/2} = v_{N/2} = w^*_{N/2}$. When all of (a), (b), and (c) occur, the values of $y_t$ generated will therefore be the same as would be generated by the basic circular coupling procedure.

From the coupling inequality (Lindvall 1992), we can bound the total variation distance between the distribution of $y_0, \ldots, y_{N-1}$ as produced by the practical procedure and the distribution of $y_0, \ldots, y_{N-1}$ as produced by the theoretical procedure by the probability of these sequences differing, which we have seen is in turn bounded above by $2\epsilon + \delta$. The total variation distance between the distributions of any individual $y_t$ for the two procedures is bounded by the same quantity. Since the distribution of each $y_t$ produced by the theoretical procedure is $\pi$, we see that the distribution of each $y_t$ produced by the practical procedure is within $2\epsilon + \delta$ of $\pi$ in total variation distance.

## 4 Testing the conditions for approximate correctness

Conditions (1) and (2) required for the approximate correctness of the circular coupling procedure will seldom be verifiable theoretically. Instead, we will have to content ourselves with an empirical diagnostic test.

This test starts by tentatively assuming that the two conditions are true for the value of $N$ we are using, and for some fairly small values of $\epsilon$ and $\delta$. If so, the value of each $y_t$ obtained should come from close to the equilibrium distribution $\pi$. Although $y_t$ will not be completely independent of $u_t, u_{t+1}, u_{t+2}, \ldots$, the dependence should in practice be sufficiently slight that we can see $y_t, y_{t+1}, y_{t+2}, \ldots$ as a realization of the Markov chain started at equilibrium, at least as long as we look only up to $y_{t+k}$ with $k \ll N$. (Here and below, addition and subtraction on times is done modulo $N$, so that if $t = N-1$, then $y_{t+1}$ refers to $y_0$.)

Condition (2) states that such a sequence, $y_t, y_{t+1}, y_{t+2}, \ldots$, will with high probability coalesce within $N$ iterations with another coupled chain started in a state drawn from the initial state distribution $p_0$. Choosing some $r$ that divides $N$, we can test whether this in fact occurs when starting at times $t = N/r$, $2N/r$, $\ldots$, $(r-1)N/r$ by simulating auxiliary chains starting at those times. In practice, we would usually wish for coalescence to occur in considerably fewer than $N$ iterations, so let us suppose that we simulate each such chain for only some number $k < N/2$ iterations, or until the auxiliary chain coalesces with $y_t, y_{t+1}, y_{t+2}, \ldots$

This leads to the following extension of the basic circular coupling procedure:

**Circularly-coupled Markov chain simulation with auxiliary diagnostic chains:**

1-5) Perform steps (1) to (5) of the basic circularly-coupled Markov chain simulation procedure.

6) Let $c_0$ be the number of steps needed for the wrapped-around chain to coalesce with the original chain — ie, let $c_0$ be the smallest $t$ such that $y_t = x_t$ — unless the chains do not coalesce within $k$ iterations, in which case let $c_0 = k$.

7) For $i = 1, \ldots, r-1$:

    Let $s = iN/r$.

    Randomly draw $z_{i,s}$ from the initial state distribution, $p_0$.

    Set $c_i = 0$.

    For $t = s+1, \ldots, s+k$ (modulo $N$) while $z_{i,t-1} \neq y_{t-1}$:
        Let $z_{i,t} = \phi(z_{i,t-1}, u_{t-1})$.
        Set $c_i = c_i + 1$.

The time required for this procedure will be roughly proportional to the number of Markov chain iterations (ie, the number of evaluations of $\phi$), which will be $N + \sum_i c_i$. Note that if all the auxiliary chains coalesce with the wrapped-around chain, $y_0, \ldots, y_N$, there will be no real distinction between the auxiliary chains and the original chain that was started with $x_0$ drawn from $p_0$. The same wrapped-around chain would have been found from any of the $r$ starting points.

The values of $c_i$ for $i = 0, \ldots, r-1$ that are obtained by this procedure are indicative of whether Condition (2) holds. If many of the $c_i$ are greater than $k$, we should not be confident that this condition holds, and should rerun the simulation with a larger value for $k$ and probably a larger value for $N$ as well (recall that $k$ should be substantially smaller than $N$). It may often be reasonable to think that the $c_i$ have an approximately geometric distribution, in which case the parameter of this distribution could be estimated from this (right-censored) data, and used to estimate the value of $\delta$ for which Condition (2) holds.

This test does not provide direct information about Condition (1), which involves two chains started from the equilibrium distribution. However, if the evidence from the auxiliary chains leads one to conclude that all but a small fraction, $q$, of chains started from $p_0$ coalesce in no more than $N/2$ iterations with a chain started from the equilibrium distribution, then one can also conclude that two chains started from the equilibrium distribution will coalesce with each other within $N/2$ iterations with probability at least $1 - 2q$, since if they both coalesce with a chain started from $p_0$, they must also coalesce with each other.

If all the auxiliary diagnostic chains are observed to coalesce quickly with the wrapped-around chain, we therefore have reason to believe that both conditions for approximate correctness hold. This will not be an absolute guarantee, however. It could be that the initial state distribution, $p_0$, gives little probability to a region that has high probability under $\pi$, and that is isolated from the regions that do have high probability under $p_0$. Both the wrapped-around chain and the auxiliary diagnostic chains might never visit this isolated region, in which case the diagnostic chains would present a self-consistent but drastically incorrect picture of the distribution of coalescence times. To help avoid this, it is desirable for $p_0$ to be "overdispersed" with respect to $\pi$, but even if this is so, there is no guarantee that all high probability regions of $\pi$ will be found, since some region with a large probability under $\pi$ might have a small "basin of attraction", and hence could be missed even if it is within the high probability region of $p_0$.

# 5 Parallel simulation

The auxiliary chains used as diagnostics in the previous section are expected to coalesce with the wrapped-around chain reasonably rapidly. If this is indeed so, coalescence of each auxiliary chain with the next auxiliary chain (started $N/r$ time steps forward) will also be fairly rapid. It will then be possible to find the wrapped-around chain by parallel computation on several processors, in less time than would be needed to simulate the wrapped-around chain using a single processor. The following procedure is based on this idea:

**Parallel simulation of a circularly-coupled Markov chain:**

In parallel, processors numbered by $i = 0, \ldots, r-1$ do the following:
(The variables $s$, $t$, and $z$ are local to each processor)

1) Let $s = iN/r$.

2) For $t = s, \ldots, s + N/r - 1$:
    Randomly draw $u_t$ from the distribution $U$, independently of other draws.

3) Randomly draw $y_s$ from the distribution $p_0$, independently of other draws.

4) For $t = s + 1, \ldots, s + N/r - 1$:
    Set $y_t = \phi(y_{t-1}, u_{t-1})$.

5) Set $z = \phi(y_{s+N/r-1}, u_{s+N/r-1})$.

6) Send $z$ to processor $i+1$ (modulo $r$) as the new value for $y_{s+N/r}$ .

7) Repeat the following:
    Wait for a new value for $y_s$ to be received from processor $i-1$ (modulo $r$).
    For $t = s + 1, \ldots, s + N/r - 1$ while $y_t \neq \phi(y_{t-1}, u_{t-1})$:
        Set $y_t = \phi(y_{t-1}, u_{t-1})$.
    If $z \neq \phi(y_{s+N/r-1}, u_{s+N/r-1})$:
        Set $z = \phi(y_{s+N/r-1}, u_{s+N/r-1})$.
        Send $z$ to processor $i + 1$ (modulo $r$) as the new value for $y_{s+N/r}$ .

The procedure terminates when all processors are waiting.

The wrapped-around chain consists of the final values of $y_0, \ldots, y_{N-1}$ that are stored in the $r$ processors. The subscript of $y$ wraps around in the above procedure, so that processor $r-1$ sends $y_N$ to processor 0, which uses it to replace the old value of $y_0$. An ordinary, non-circular Markov chain simulation can be performed in parallel in the same way as above by omitting this wrap-around, keeping $y_0$ fixed at its original value, but I will not discuss this possibility further here.

The computation time for the above simulation will be at least the time required to simulate $N/r$ Markov chain iterations, since that many iterations will always be done in steps (4) and (5). Each processor will then begin simulating a chain starting from the new value received for its $y_s$. If each of these chains coalesces within $N/r$ iterations with the chains that were simulated starting from the original values for each $y_s$, then each processor will find that the value for $y_{s+N/r}$ that it communicates to the next processor is unchanged, and the entire procedure will terminate. The time taken will be that required for between $N/r$ and $2N/r$ Markov chain iterations.

If, on the contrary, not all of these chains coalesce within $N/r$ iterations, one or more of the processors will have to perform a third simulation. In general, a processor might have to rerun its simulation any number of times, as a result of the previous processor sending it new start states. Assuming that a sequential simulation would have resulted in the wrapped-around chain coalescing with the original chain, the time required for the parallel simulation will be roughly proportional to the maximum number of iterations that any of the $r$ chains with different starting points take to coalesce with this wrapped-around chain. If this time is comparable to the time for a sequential simulation, the slow coalescence of the auxiliary chain would be indicative of a problem, and it would usually be best to stop the whole procedure, and restart it with a larger value for $N$.

It is possible that the procedure as shown will not terminate. This will happen if different starting points lead to different wrapped-around chains; Figure 6 in Section 7 below illustrates this possibility. In practice, the procedure should be terminated when some processor has received more than some maximum of new values for its starting point. The maximum number of such new starting points is a rough diagnostic of how rapidly the chains couple, providing information similar to that provided by the $c_i$ in the procedure of Section 4.

This parallel simulation procedure may be adaptable to vector computation, provided the computation of $\phi$ does not involve lots of conditional computations. Such a vectorized simulation might be appropriate when vector operations are supported by hardware, or when programming is done in an interpreted language in which vector operations are not much slower than scalar operations, due to the fixed overhead of interpretation.

Finally, one should note that although the parallel simulation procedure aims to do roughly what is done by the sequential procedure with auxiliary diagnostic chains of Section 4, the actual computations done may differ. When not all chains coalesce with the wrapped-around chain within $N/r$ iterations, the sequential procedure of Section 4 will simulate two or more chains that operate at the same times. It is possible that these chains will coalesce with each other before coalescing with the wrapped-around chain, but the sequential procedure does not detect this, and will simulate the coalesced chains separately. Coalescence of chains is detected differently in the parallel procedure, however, which may lead to such portions of chains being simulated only once. It is therefore conceivable that it would be advantageous to use the parallel procedure even when one has available only a single processor, which could execute the $r$ parallel processes of the algorithm by time sharing.

## 6 A Metropolis coupling scheme

To be practically useful, the circular coupling procedure needs a representation of the Markov chain transitions in terms of a $\phi$ function that is easily computable, and that leads to rapid coalescence of chains. Exact sampling methods such as coupling from the past (Propp and Wilson 1996) also require such a representation, as do the coupling-based schemes of Johnson (1996, 1998). Several such schemes have recently been discussed in the context of exact sampling by Green and Murdoch (1998). Note that one is at liberty to choose the Markov chain transitions to facilitate coupling, though this might sometimes come at a cost in terms of convergence rate.

I will not discuss coupling schemes in any depth in this paper. Instead, I will present a simple coupling scheme for Metropolis updates in continuous state spaces that leads to exact coalescence with positive probability, which I will use in the demonstrations in the next section. On its own, this scheme will not be adequate for all real problems, but it may be generally useful in producing exact coalescence once other methods have brought the chains to nearby states.

Suppose that our state, $x$, consists of a single real value, and that our desired distribution is given by the density function $\pi(x)$. Recall that the Metropolis algorithm defines a Markov chain transition in terms of a density function, $g(x^*|x)$, for proposing a move to state $x^*$, given that the chain is currently in state $x$. This proposal is accepted with probability $\min[1, \pi(x^*)/\pi(x)]$. If the proposal is rejected, the new state is the same as the old state. Provided that the proposal distribution is symmetrical (ie, $g(x^*|x) = g(x|x^*)$), this update leaves the distribution $\pi$ invariant.

I will consider a random-walk Metropolis algorithm using proposals that are uniformly distributed in an interval of width $w$, centred on the current state, for which

$$g(x^*|x) \;=\; \begin{cases} 1/w & \text{if } |x^* - x| < w/2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The most obvious way of expressing this update in terms of a function $\phi$ is as follows:

$$\phi(x,u) \;=\; \begin{cases} x + w(u_1 - 1/2) & \text{if } u_2 < \pi(x + w(u_1 - 1/2))\,/\,\pi(x) \\ x & \text{otherwise} \end{cases} \tag{2}$$

This function takes the current state and a vector of two Uniform$(0,1)$ random numbers as arguments, and returns the next state, which will have the distribution as defined for the Metropolis algorithm with the proposal distribution above. However, with this $\phi$ function, the probability of exact coalescence of chains is zero, whenever the current states are distinct.

Fortunately, the same Markov chain transition probabilities are obtained with the following $\phi$ function, which can lead to coalescence with positive probability:

$$\phi(x,u) \;=\; \begin{cases} f(x,u) & \text{if } u_2 < \pi(f(x,u))\,/\,\pi(x) \\ x & \text{otherwise} \end{cases} \tag{3}$$

$$\text{where } f(x,u) \;=\; w\,[(u_1 - 1/2) + \text{Round}(x/w - (u_1 - 1/2))]$$

Round returns the integer nearest its argument. The function $f(x,u)$ can be seen as first transforming the state to $x' = x/w - (u_1 - 1/2)$, then rounding to the nearest integer, then applying the inverse transformation. Clearly, for a given value of $u_1$, a range of values for $x$ of width $w$ will all result in the same value for $f(x,u)$. Two chains whose current states are in this range, and for which the proposed point is accepted, will therefore coalesce exactly as a result of this transition.

This $\phi$ function can be visualized as first laying down a grid of points spaced $w$ apart, with the position of the grid being chosen uniformly at random, and then proposing to move to the point on this grid that is nearest the current state, $x$. It is then clear that the distribution of the proposed state is as in equation (1). The scheme generalizes directly to higher dimensions, in which the proposal distribution is uniform over a hypercube centred on the current state, by simply using a higher-dimensional grid.

# 7 Simple demonstrations

To illustrate the concept of circularly-coupled Markov chain simulation, I include here two simple illustrations with a one-dimensional state space. Both examples use the random-walk Metropolis algorithm based on the uniform proposal distribution of equation (1), with $w = 1$, coupled using the scheme of equation (3).

The first example illustrates the behaviour of circular coupling when coalescence is rapid. Figure 3 shows a simulation of length $N = 1000$ that samples from the $N(0,1)$ distribution, with $N(0,5^2)$ as the initial state distribution. Ten chains were simulated in total — one started at $t = 0$, plus nine auxiliary chains started at $t = 100, 200, \ldots, 900$. All chains coalesce rapidly with the wrapped-around chain, an indication (but not an absolute guarantee) that the conditions for the states of the wrapped-around chain to all come from the equilibrium distribution are satisfied.

The second example illustrates how circular coupling behaves when coalescence is less rapid. The distribution to be sampled from is in this case a bimodal mixture of normals, $(3/4)N(-1,1) + (1/4)N(1.5, 0.1^2)$. Figure 4 shows a circularly-coupled simulation for this distribution that is typical of runs of length $N = 1000$, with $r = 10$ starting points. Some of the ten chains started with states from the $N(0,5^2)$ distribution coalesce rapidly, as for the first example, but others do not. In particular, the chain started at $t = 900$ takes about 400 iterations to coalesce with the wrapped-around chain. This is evidence that the conditions needed for the states of the wrapped-around chain to come from approximately the correct distribution may not be satisfied.

Figure 5 shows another run of the same simulation, with a different pseudo-random number seed. This run might well be misleading, since the wrapped-around chain found samples from only the lower of the mixture distribution's two modes. Moreover, the chains from all ten starting points coalesce with this wrapped-around chain reasonably quickly, which might lead one to think that the procedure is sampling from approximately the correct distribution. The distribution of states produced by this procedure might indeed be at least roughly correct, since as seen in the previous figure, other runs of the same simulation procedure do produce wrapped-around chains that visit the upper mode as well. However, even if the distribution of each state of the wrapped-around chain for $N = 1000$ is close to correct, the states in this chain are clearly highly dependent, and hence a single wrapped-around chain can fail to provide an adequate sample. Behaviour similar to this occurs in a few percent of the runs.

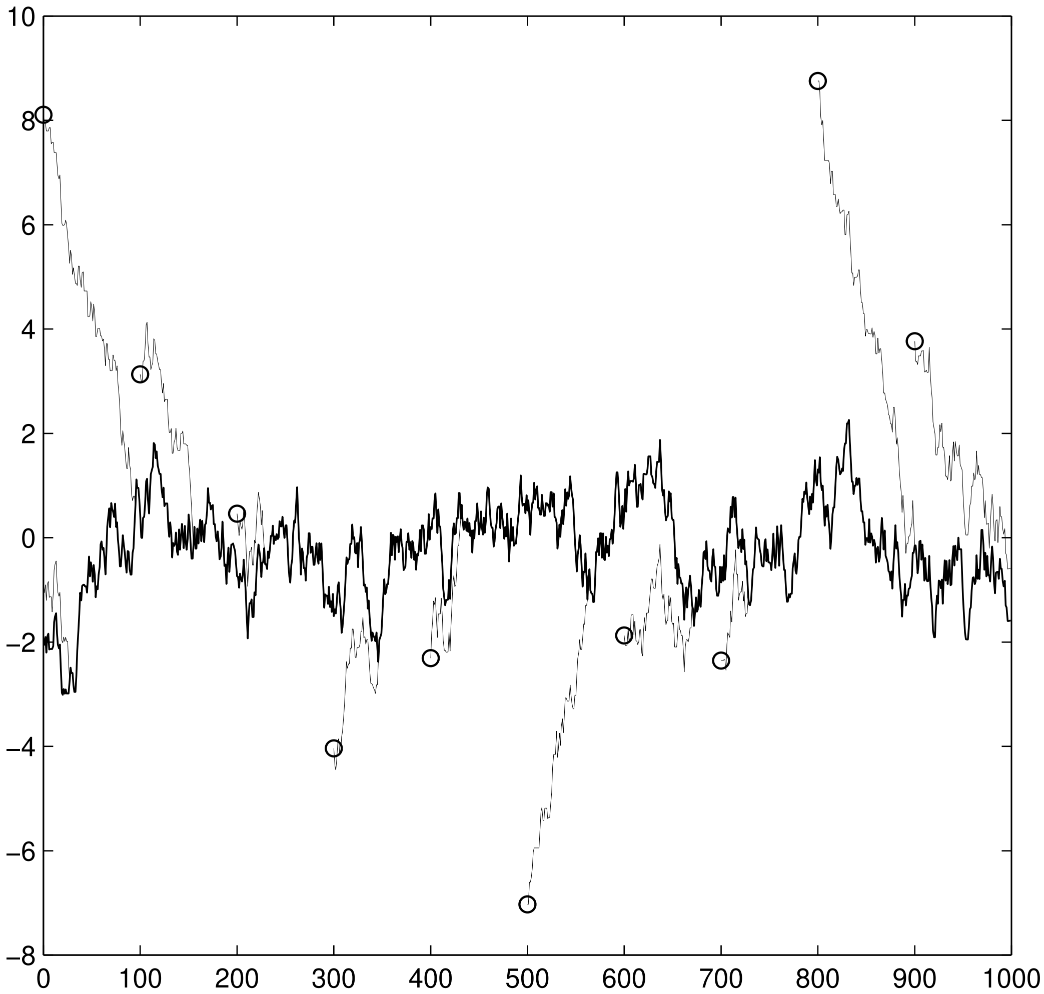


Figure 3: A circularly-coupled random-walk Metropolis simulation sampling from the $N(0,1)$ distribution. Ten chains started with states drawn from the initial distribution $N(0,5^2)$ (shown as circles) all coalesce with the wrapped-around chain (the thick line) in less than 150 iterations, much less than the total of 1000. This is consistent with the conditions required for the circular coupling procedure to be approximately correct.

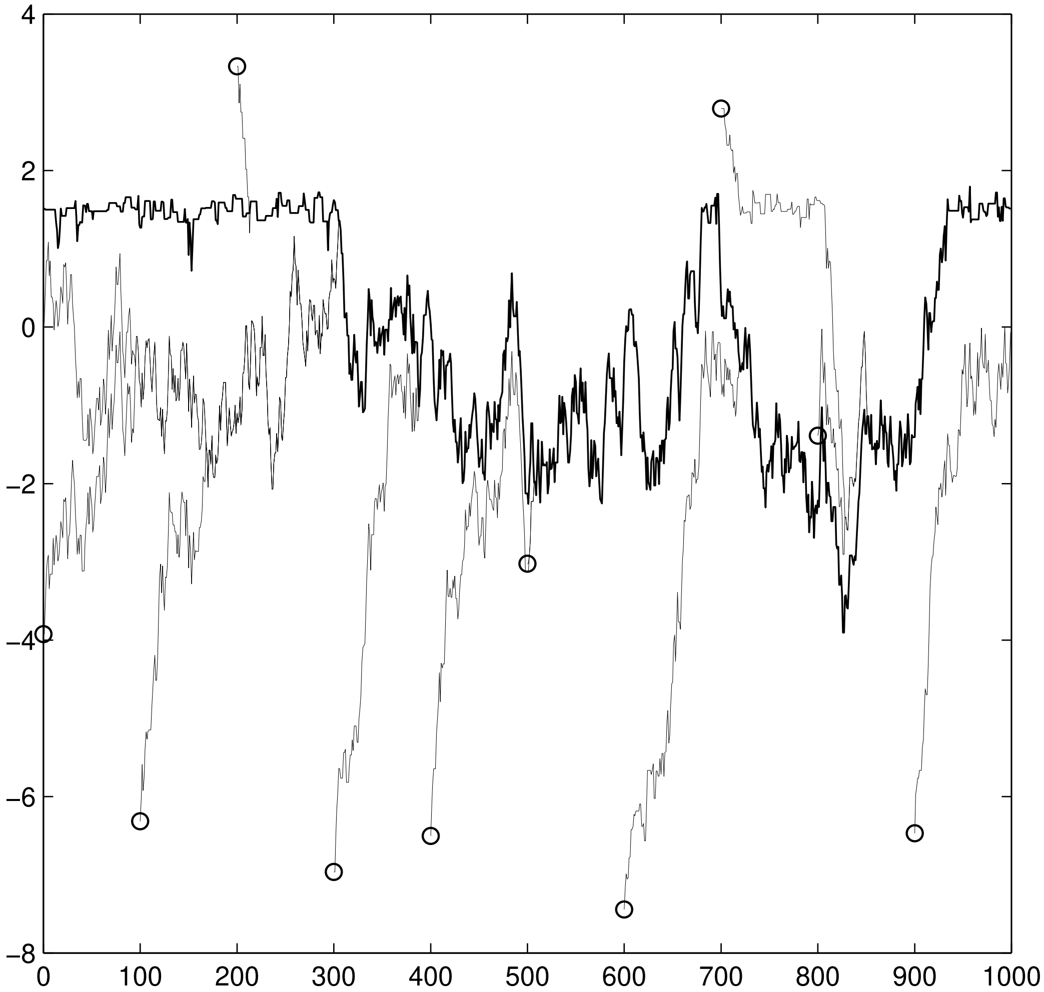


Figure 4: A circularly-coupled random-walk Metropolis simulation sampling from the distribution $(3/4)N(-1,1) + (1/4)N(1.5, 0.1^2)$. One of the ten chains started from the initial distribution $N(0, 5^2)$ takes 400 iterations to coalesce with the wrapped-around chain. Since this is a substantial fraction of the total of 1000 iterations, one might doubt whether the conditions for the circular coupling procedure to be approximately correct are satisfied.

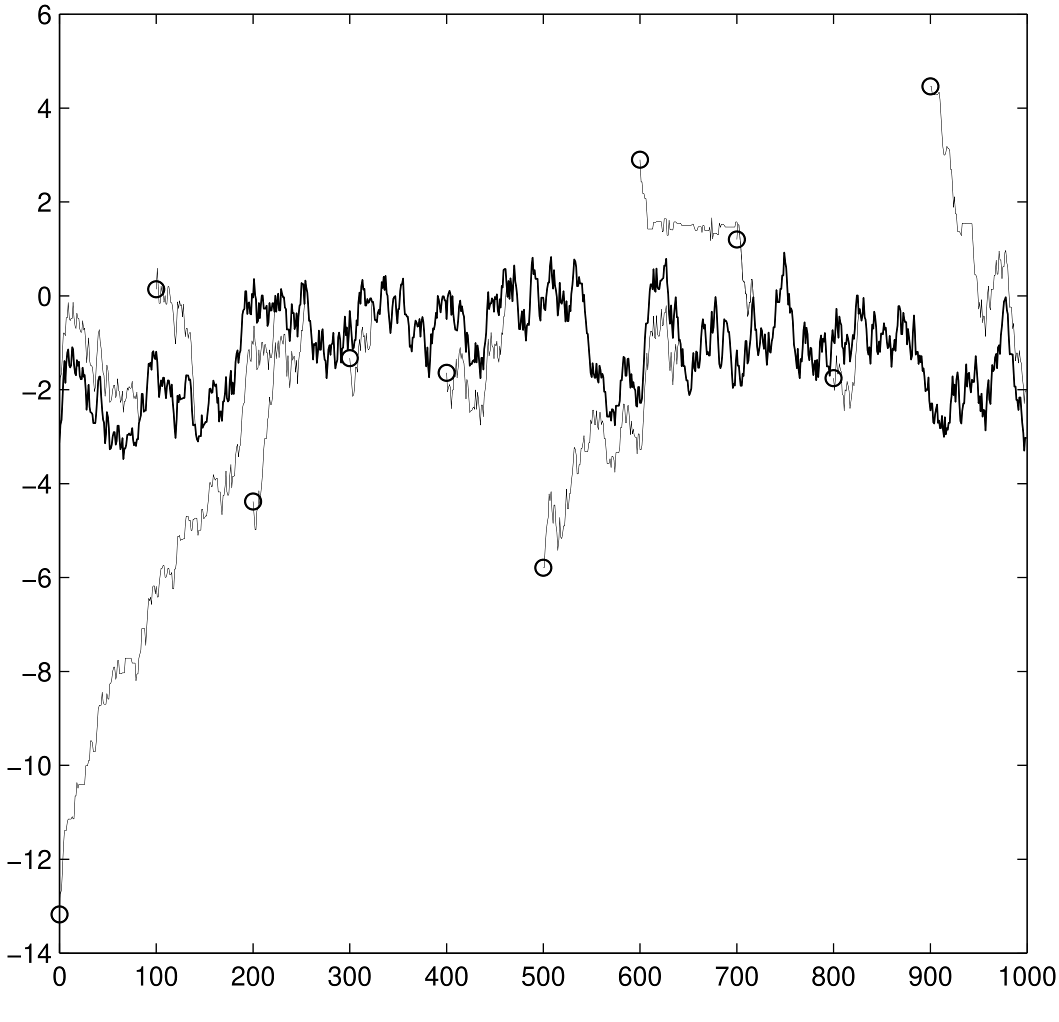


Figure 5: Another run of the circularly-coupled simulation shown in Figure 4. This time the ten chains all coalesce with a wrapped-around chain that has visited only the lower of the two modes.

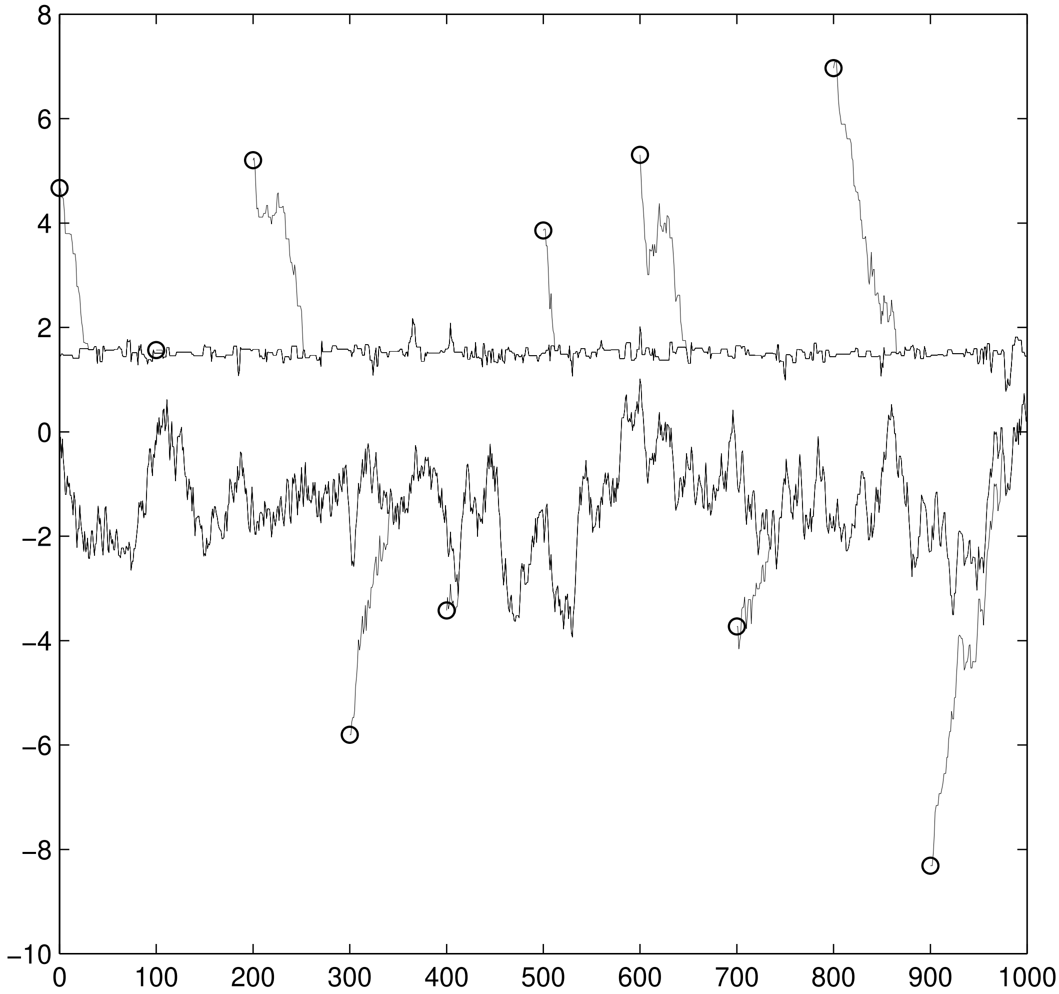


Figure 6: A run of the same circularly-coupled simulation as in Figure 4 in which a single wrapped-around chain was not found after simulating each chain for $N$ iterations. Instead, six of the chains coalesced into a wrapped-around chain that stays in the upper mode, while the other four chains coalesced into a wrapped-around chain that stays in the lower mode.

Using more than ten random starting points would reduce the chances of such a problem remaining undiagnosed.

Figure 6 shows another possible result of the simulation procedure, which occurs about once in a thousand runs. Here, six of the ten chains coalesce to a wrapped-around chain that moves within the upper mode only. The chains started from the other four initial states coalesce to a wrapped-around chain that moves within the lower mode. When this situation occurs, one can tell that the simulation should be rerun with a larger value of $N$.

## 8 Discussion

The circular coupling procedure has been shown here to produce states with approximately the correct distribution provided certain conditions regarding coalescence times are satisfied. In practice, we will usually not know with certainty whether these conditions hold, but diagnostic tests can provide useful evidence of this.

If we did have a theoretical proof that the required conditions hold, the benefits of circular coupling would be modest. These conditions would also suffice to show that the distribution of the last state of an ordinary Markov chain simulation for $N$ time steps comes from close to the equilibrium distribution of the chain. We could therefore obtain a sample of $N$ (dependent) points from approximately the correct distribution by simply continuing the simulation of this chain for another $N$ iterations. The circular coupling procedure would save at most a factor of two in computation time compared to this alternative, less if coalescence of the wrapped-around chain did not occur quickly.

The primary reason why circular coupling is of interest is that it may provide a way of diagnosing convergence and discarding a "burn-in" period that is more automatic than current methods, allowing Markov chain Monte Carlo methods to be used on a more routine basis. The ability of circular coupling to exploit parallel computation may also be useful in practice. To obtain these benefits, effective coupling schemes will need to be developed, which can be applied to a wide range of distributions with little or no need for problem-specific tailoring.

## Acknowledgements

I thank Jeffrey Rosenthal for inspiration to examine the possibility of parallel simulation of Markov chains and for helpful discussions. This research was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Institute for Robotics and Intelligent Systems.

## References

Brooks, S. P. and Roberts, G. O. (1998) "Convergence assessment techniques for Markov chain Monte Carlo", *Statistics and Computing*, vol. 8, pp. 319-335.

Cowles, M. K. and Carlin, B. P. (1996) "Markov chain Monte Carlo convergence diagnostics: a comparative study", *Journal of the American Statistical Association*, vol. 91, pp. 883-904.

Fill, J. A. (1998) "An interruptible algorithm for perfect sampling via Markov chains", *Annals of Applied Probability* vol. 8, pp. 131-162.

Frenkel, D. and Smit, B. (1996) *Understanding Molecular Simulation: From Algorithms to Applications*, San Diego: Academic Press.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Green, P. J. and Murdoch, D. J. (1998) "Exact sampling for Bayesian inference: towards general purpose algorithms" (with discussion), in J. M. Bernardo, *et al* (editors), *Bayesian Statistics 6*, Oxford: Clarendon Press, pp. 301-321.

Johnson, V. E. (1996) "Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths", *Journal of the American Statistical Association*, vol. 91, pp. 154-166.

Johnson, V. E. (1998) "A Coupling-Regeneration Scheme for Diagnosing Convergence in Markov Chain Monte Carlo Algorithms", *Journal of the American Statistical Association.*

Lindvall, T. (1992) *Lectures on the Coupling Method*, New York: Wiley.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999) "MCMC convergence diagnostics: a reviewww" (with discussion), in J. M. Bernardo, *et al* (editors), *Bayesian Statistics 6*, Oxford: Clarendon Press, pp. 415-440.

Propp, J. G. and Wilson, D. B. (1996) "Exact sampling with coupled Markov chains and applications to statistical mechanics", *Random Structures and Algorithms*, vol. 9, pp. 223-252.

Rosenthal, J. S. (1995a) "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo", *Journal of the American Statistical Association*, vol. 90, pp. 558-566.

Rosenthal, J. S. (1995b) "Convergence rates of Markov chains", *SIAM Review*, vol. 37, pp. 387-405.

Rosenthal, J. S. (1999) "Parallel computing and Monte Carlo algorithms", Technical Report No. 9902, Dept. of Statistics, University of Toronto.

Sinclair, A. (1993) *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Boston: Birkhäuser.