CS294 Markov Chain Monte Carlo: Foundations & Applications **Fall 2009**

Lecture 9: October 6

Lecturer: Prof. Alistair Sinclair *Scribes: Alistair Sinclair*

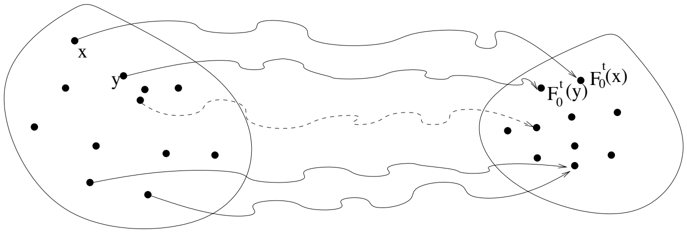
Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

9.1 Random function representation

Let P be an ergodic Markov chain on state space Ω. Recall the "random function" representation of P, namely as a probability distribution over functions $f : \Omega \to \Omega$ such that

$$\Pr[f(x) = y] = P(x, y) \qquad \forall x, y, \tag{9.1}$$

where the probability is over the random choice of f.



Let us fix a labeling of steps as "points in time" $(-\infty, \ldots, -1, 0, 1, \ldots, \infty)$ with arbitrary "current time" 0. Define F_i^j as the $(j-i)$-step evolution of the Markov chain M from time i to j, and decompose it into $(j-i)$ independent applications of a random function f consistent with M; that is

$$F_i^j = f_{j-1} \circ f_{j-2} \circ \cdots \circ f_{i+1} \circ f_i \tag{9.2}$$

where f_t is the random function chosen at time t. Two cases will play a special role for us in this lecture: firstly, F_0^t is the standard "forwards" simulation of M for t steps (starting at time 0). Secondly, F_{-t}^0 is the t-step evolution of M from time $-t$ to time 0 (i.e., "from the past")

9.2 Coupling from the past

Define the "coalescence time" T as

$$T = \min\{t : F_0^t \text{ is a constant function}\}. \tag{9.3}$$

Note that F_0^t being constant means that, after t steps, all paths in the Markov chain have reached the same state (for all possible initial states).

In view of what we know about coupling, we might be tempted to conjecture that the distribution of $F_0^T(x)$ (which, by definition, is the same for all x) is the stationary distribution π. However, as we shall see in a moment, this is spectacularly false. But remarkably, if we try the same trick “from the past”, we *do* get the correct distribution.

To state this formally, define

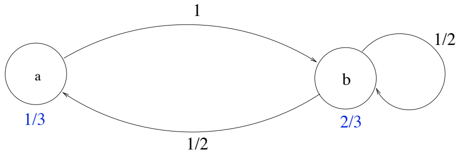
$$T = \min\{t : F^0_{-t} \text{ is a constant function}\}. \tag{9.4}$$

The following theorem, due to Propp and Wilson [PW96], shows that if we take T as a stopping time for the simulation from the past, then the resulting constant value of the function F^0_{-T} will have distribution exactly π:

Theorem 9.1 *Assuming T is finite with probability 1, then the constant value $Z^0_{-\infty} = F^0_{-T}(x)$ has distribution exactly π.*

Before proving the theorem, we will give an example to illustrate the difference between the forward and “from the past” simulations.

Consider the following very simple Markov chain M, whose stationary distribution π is $(1/3, 2/3)$:



In this case, there is a *unique* random function f consistent with M, as follows:

$$\begin{array}{ll} f_1(a) = b & f_2(a) = b \\ f_1(b) = a & f_2(b) = b \end{array}$$

$$f = \begin{cases} f_1 \text{ with probability } 1/2 \\ f_2 \text{ with probability } 1/2 \end{cases}$$

Exercise: Verify that f is unique.

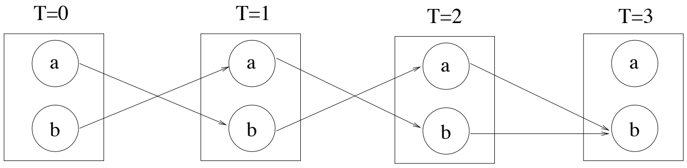


Figure 9.1: One particular forward simulation where $T = 3$. Note that $F^T_0(x) = b$ always.

Figure 9.1 shows a sample forwards simulation with stopping time $T = 3$. You should convince yourself that such a simulation must always stop in state b, i.e., in distribution $(0, 1)$, which is very far from the stationary distribution $(1/3, 2/3)$. Figures 9.2 and 9.3 show two possible simulations from the past, one of which stops in state b and the other in state a.

Exercise: Verify that in the above example, the distribution of F^0_{-T} is indeed π.

Exercise: Show that, although the distributions of F^0_{-T} and F^T_0 (where in each case T is the appropriate stopping time) are quite different, the distributions of the two *stopping times* are the same.

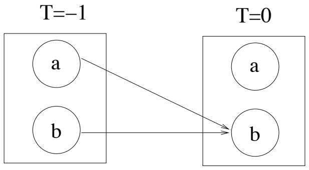


Figure 9.2: Simulation from the past with $T = 1$.

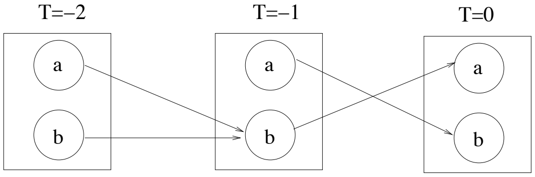


Figure 9.3: Simulation from the past with $T = 2$.

Now we will go back and prove Theorem 9.1.

Proof: Since T is finite with probability 1, $Z^0_{-\infty}$ is well defined with probability 1. Define

$$T' = \min\{t : F^1_{-t} \text{ is constant}\}.$$

Couple the processes for F^0_{-t} and F^1_{-t}, so that they use the same $f_{t'}$ at time t' (hence $T' \leq T$). Let $Z^1_{-\infty} = F^1_{-T'}(x)$ be the constant value of the function $F^1_{-T'}$, and π_0, π_1 be the distributions of $Z^0_{-\infty}$ and $Z^1_{-\infty}$ respectively. Since $Z^0_{-\infty}$, $Z^1_{-\infty}$ have the same distribution (both are just the value of the constant function obtained by coupling from the past up to some fixed time, 0 or 1 respectively), we have $\pi_0 = \pi_1$. Also, it should be clear that $Z^1_{-\infty} = f(Z^0_{-\infty})$, since the former is obtained by extending the simulation for one further step. Hence the common distribution $\pi_0 = \pi_1$ is a fixed point of f, so it must be the (unique) stationary distribution of M. ■

9.3 An exact sampling algorithm

Theorem 9.1 immediately suggests an algorithm for sampling *exactly* from the stationary distribution π. (Note that our standard Markov chain simulation does not quite achieve this: there is always some non-zero variation distance.) Here is the algorithm:

1. Compute F^0_{-1}, F^0_{-2}, …, F^0_{-t}, … until F^0_{-t} is constant.
2. Output the constant value of F^0_{-t}.

Unfortunately this algorithm is in general not very efficiently implementable: in order to check if F^0_{-t} is constant, we need to compute $F^0_{-t}(x)$ for every state x. This is hopeless in most of our examples, where the number of states $|\Omega|$ is huge.

However, there is a situation in which we can turn the above into an efficient algorithm: namely, if the random function f is a *monotone* coupling then to test if F^0_{-t} is constant it is enough just to test if $F^0_{-t}(\top) = F^0_{-t}(\bot)$, where $\top$ and $\bot$ are the unique maximum and minimum elements respectively (see Figure 9.4). [**Exercise:** Check carefully that this is equivalent to testing if F^0_{-t} is constant.]

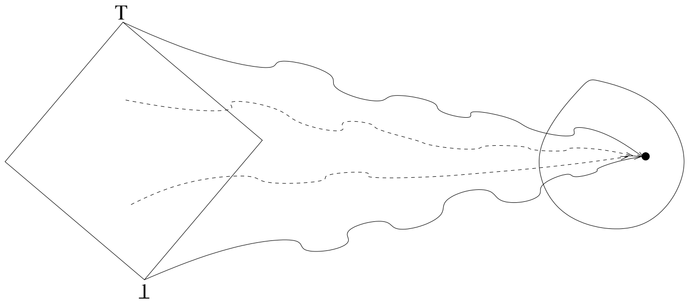


Figure 9.4: For a monotone coupling, top and bottom states converging to the same state after time T implies that all other states converge to the same state by that time.

We can gain a further improvement in running time by noting that, instead of trying $F^0_{-1}, F^0_{-2}, F^0_{-3}, F^0_{-4}, \ldots$ in sequence, it is enough to try the times $F^0_{-1}, F^0_{-2}, F^0_{-4}, F^0_{-8}, \ldots$. This is OK because "overshooting" the stopping time T does not affect the distribution of the final state. (I.e., if F^0_{-t} is constant and has distribution π, then the same is true of $F^0_{-t'}$ for all $t' \geq t$.) And if the true stopping time is T, then the above doubling scheme will stop after at most $2T$ steps. Moreover, the number of steps it will simulate is at most $1 + 2 + 4 + 8 + \cdots + 2T \leq 4T$. Contrast this with the naive single-step scheme which will simulate $1 + 2 + 3 + 4 + \cdots + T = \Omega(T^2)$ steps.

Here, then, is the final algorithm for the monotone case:

```
$T \leftarrow 1$
**repeat**
    bottom $\leftarrow \perp$
    top $\leftarrow \top$
    **for** $t \leftarrow -T$ **to** $-1$ **do**
        bottom $\leftarrow f_t$(bottom)
        top $\leftarrow f_t$(top)
        $T \leftarrow 2T$
    **end**
**until** bottom=top ;
**Output**: top
```

Algorithm 1: A "Coupling from the Past" algorithm for exact sampling from π

Note that, in this algorithm, it is important to store and re-use the random choices f_t as you go, rather than to recompute them at each step. [**Why?**]

It is interesting to observe that, in the monotone setting, Coupling from the Past never adds a huge overhead compared to standard forwards simulation of the Markov chain: specifically, one can show that the expected value of the coalescence time T from the past is bounded above by $O(h\tau_{\text{mix}})$, where h is the *height* of the partial order (i.e., the length of a longest chain of comparable elements between $\perp$ and $\top$). [**Exercise**: Prove this!] Thus in order to bound *a priori* the amount of time required to produce a sample, it is enough to bound the mixing time. In many practical situations, however, the actual time until Coupling from the Past produces a sample is much less than the (best known estimate of) the mixing time. Note also that we can use CFTP even when no analytical estimate of the running time is available (but then we have no bound on how long we may have to wait for a result).

Finally, we should inject a caveat here. Coupling from the Past only guarantees a sample from distribution π if we are willing to wait as long as it takes for the procedure to terminate. If, e.g., we truncate (and restart) the algorithm whenever it runs for too long, we can introduce an arbitrary bias into the distribution. There are so-called "interruptible" versions of CFTP (see, e.g., [Fi98,FMMR00]) which do allow one to terminate without introducing bias, but these are much less elegant and less useful in practice than the basic version above.

9.4 Extensions

The condition of monotonicity is in fact not necessary for an efficient implementation of CFTP; in certain non-monotone settings the algorithm can be made to work efficiently also. Recall that the essential issue is testing efficiently whether the random function F_{-t}^0 is constant. Here is a trick, due to Häggstrom and Nelander [HN98] and Huber [Hu98], for doing this in the case of the Hard Core Model (or Independent Sets), one of the most important models in Statistical Physics.

Here we are given a graph $G = (V, E)$ and a parameter $\lambda > 0$, and the set of configurations Ω is the set of all independent sets (of any size) in G. We refer to vertices in the independent set as "occupied", and the remainder as "unoccupied". Thus the constraint is that no two adjacent vertices may be occupied. Each independent set $x \in \Omega$ has weight $w(x) = \lambda^{|x|}$, and as usual our goal is to sample from the distribution $\pi(x) \propto w(x)$.

The heat-bath dynamics for this model makes transitions from configuration x as follows:

- pick a vertex $v \in V$ u.a.r. (and ignore the current state of v)
- with probability $\frac{1}{1+\lambda}$ make v unoccupied; with probability $\frac{\lambda}{1+\lambda}$ make v occupied if possible (i.e., if none of its neighbors is occupied), else make it unoccupied.

Exercise: Verify that this Markov chain is ergodic and reversible w.r.t. the distribution π.

The natural random function (complete coupling) representation of this dynamics is the following. We pick a pair (v, r) u.a.r., where $v \in V$ and $r \in [0, 1]$; then we make the update at vertex v, using r to decide whether to try to occupy v or not (i.e., if $r \leq \frac{1}{1+\lambda}$ make v unoccupied, else make it occupied if possible).

There is no monotone coupling for this dynamics, so the standard approach to CFTP of the previous section does not work. (Of course, we can always apply CFTP to any Markov chain; the problem is that we have no efficient way of telling when F_{-t}^0 is constant.) The trick is to consider a modified Markov chain in which each vertex may be either occupied, unoccupied, or in state "?". The idea is that any configuration in this enlarged model represents a *set* of configurations in the original model: a vertex that is (un)occupied in the enlarged model means that that vertex is (un)occupied in every configuration in the set; a vertex that is "?" in the enlarged model means that that vertex may be both occupied and unoccupied in configurations in the set.

Transitions in the enlarged model are made as follows:

- pick a vertex $v \in V$ u.a.r. (and ignore the current state of v)
- with probability $\frac{1}{1+\lambda}$ make v unoccupied; with probability $\frac{\lambda}{1+\lambda}$, make v occupied if all of its neighbors are unoccupied, make it unoccupied if all of its neighbors are occupied, else make it "?".

Thus after each such update the 0, 1, ? configuration describes the set of possible independent sets that the original Markov chain could be in after one of the original heat-bath updates. This dynamics has essentially the same random function representation as above.

Now it is not hard to check **[exercise!]** that if the enlarged chain, starting from the all-? configuration, ever reaches a configuration y in which there are no ?'s, then the original heat-bath dynamics on independent sets, using the same realizations of the random function at each step, maps every initial configuration to the same destination, namely this same y. Hence we can use this as an efficient test for coalescence.

We face the same challenge as with standard CFTP if we want to estimate the running time in advance (and in fact we can't appeal to the previously mentioned result relating the coalescence time to the mixing time, since that required monotonicity). It has been shown [HN99] that the number of "?"'s decays exponentially fast provided $\lambda < \frac{1}{\Delta}$, where Δ is the maximum degree of G. Thus CFTP can be implemented efficiently in this range. In practice, however, for many graphs of interest (such as regular grids) the method works efficiently for much larger values of λ than this theoretical guarantee.

We will have more to say about the Hard Core Model, and the significance of the parameter λ, later in the course.

References

[Fi98] J.A. Fill, "An interruptible algorithm for perfect sampling via Markov chains", *Annals of Applied Probability* **8** (1998), pp. 131–162.

[FMMR00] J.A. Fill, M. Machida, D.J. Murdoch and J.S. Rosenthal, "Extension of Fill's perfect rejection sampling algorithm to general chains", *Random Structures & Algorithms***17** (2000), pp. 290–316.

[HN98] O. Häggstrom and K. Nelander, "Exact sampling from anti-monotone systems", *Statistica Neerlandica* **52** (1998), pp. 360–380.

[HN99] O. Häggstrom and K. Nelander, "On exact simulation of Markov random fields using coupling from the past", *Scandinavian Journal of Statistics* **26** (1999), pp. 395–411.

[Hu98] M. Huber, "Exact sampling and approximate counting techniques", *Proceedings of the 30th STOC*, 1998, pp. 31–40.

[PW96] J. Propp and D. Wilson, "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics", *Random Structure and Algorithms*, 1996, pp. 9:223-252.