Markov Chains and Algorithmic Applications: WEEK 12

1 Exact Simulation

In the previous lectures, we saw several procedures to generate samples from a distribution π on S. Moreover, the Metropolis-Hastings procedure allows us to do so with only partial knowledge of π, but this approach is merely approximate, as we do not know how long to run the algorithm to reach the stationary distribution π. In other words, the best we can do is upperbound the mixing time T_ϵ so as to sample from a distribution that is arbitrarily close to π.

Today, we will see another procedure named *coupling from the past* (introduced by James Propp and David Wilson in 1996) which allows us to sample *exactly* from π. In order to study this method, we need a tool called *random mapping representation of a Markov chain.*

Please note that in the following, we consider an ergodic Markov chain X with finite state space S, transition matrix P, and limiting and stationary distribution π.

1.1 Random Mapping Representation

So far we used to define a (time-homogeneous) Markov chain by a matrix of transition probabilities $P = (p_{ij})$, where $p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$. Alternatively, one can represent a Markov chain as

$$X_{n+1} = \Phi(X_n, U_{n+1})$$

where $\Phi(\cdot, \cdot)$ is a cleverly chosen function and $(U_n, n \geq 1)$ is a sequence of i.i.d. random variables.

Proposition 1.1. Every Markov chain admits a random mapping representation.

Proof. We assume the U_n's are uniform random variables in $[0, 1]$ (we denote this as $U_n \sim \mathcal{U}[0, 1]$) and construct $\Phi(\cdot, \cdot)$ such that $p_{ij} = \mathbb{P}\left(X_{n+1} = j | X_n = i\right) = \mathbb{P}\left(\Phi(i, U_{n+1}) = j\right)$ for any arbitrary set of transition probabilities p_{ij}. Define now

$$F_{ik} \triangleq \sum_{j=1}^{k} p_{ij}, \qquad \forall i, k \in S$$

(where S is the state space) and set

$$\Phi(i, u) \triangleq \sum_{j \in S} j \cdot \mathbb{1}\left\{F_{i,j-1} < u \leq F_{ij}\right\}.$$

We hence have

$$\mathbb{P}(\Phi(i, U_{n+1}) = j) = \mathbb{P}(F_{i,j-1} < U_{n+1} \leq F_{ij}) = F_{ij} - F_{i,j-1} = p_{ij}. \qquad \square$$

Remark: In general, there may exist many different random mapping representations for a particular chain. In the above proof we just constructed *one of* these representations.

1.2 Forward Coupling

Suppose we take two copies of a Markov chain X_n and Y_n having stationary distribution π. Their random mapping representations are

$$X_{n+1} = \Phi(X_n, U_{n+1})$$
$$Y_{n+1} = \Phi(Y_n, U_{n+1}).$$

In general, the U_n's used in the chains X_n and Y_n are two independent samples. However, *if we use the same samples U_{n+1} for updating $X_n \to X_{n+1}$ and $Y_n \to Y_{n+1}$,* we will impose *grand coupling* between those chains.

Now suppose we start $|S|$ copies of the chain $\left(X_n^{(i)}, i \in S\right)$, each starting at a different state (i.e. $X_0^{(i)} = i$), and update them using the same samples of U_n (i.e., we establish pairwise grand coupling). This situation is called *forward coupling*.

One may think that once all chains coalesce at some time $T > 0$, the initial state has been "forgotten", so that the stationary distribution has been reached (i.e. $\forall i, j \in S, \quad \mathbb{P}\left(X_T^{(i)} = j\right) = \mathbb{P}\left(X_T = j\right) = \pi_j$). Unfortunately, this is not the case, as we will see in the following examples:

Example 1.2. Consider the Markov chain of Figure 1. A random mapping representation of this chain (using $U_n \sim \mathcal{U}[0,1]$) is

$$\Phi(0,u) = \begin{cases} 0 & \text{if } u \leq \frac{1}{2}, \\ 1 & \text{if } u > \frac{1}{2}, \end{cases}$$
$$\Phi(1,u) = 0.$$

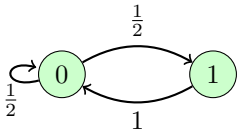


Figure 1: Markov chain of Example 1.2

It is easy to check that coalescence always happens at state 0. Indeed, the only way to get $\Phi(0,u) = \Phi(1,u)$ is to have $0 \leq u \leq \frac{1}{2}$ implying $\Phi(0,u) = \Phi(1,u) = 0$. For example, consider the situation depicted in Figure 2. That is to say, $(\pi_0(T) = 1, \pi_1(T) = 0)$ where T is the coalescence time.

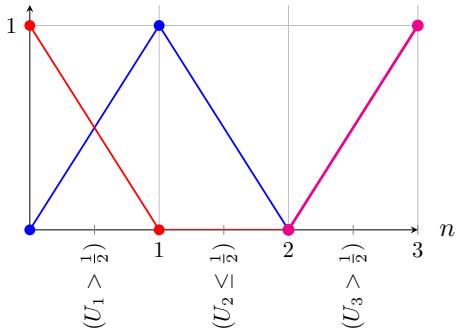


Figure 2: Two copies of the chain considered in Example 1.2.

However, the stationary distribution is $\left(\pi_0 = \frac{2}{3}, \pi_1 = \frac{1}{3}\right)$. Therefore, the chains are not in the stationary distribution when they coalesce. They are also not in the stationary distribution after the coalescence time.

The choice of the random mapping representation can even lead to situations where we do not have coalescence at all.

Example 1.3. Consider the Markov chain of Figure 3. One possible candidate for its random mapping representation (still assuming $U_n \sim \mathcal{U}[0,1]$) is

$$\Phi(i,u) = \begin{cases} i & \text{if } u \leq \frac{1}{3}, \\ 1-i & \text{if } u > \frac{1}{3} \end{cases}$$

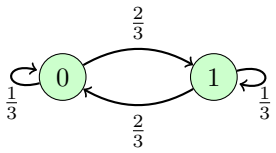
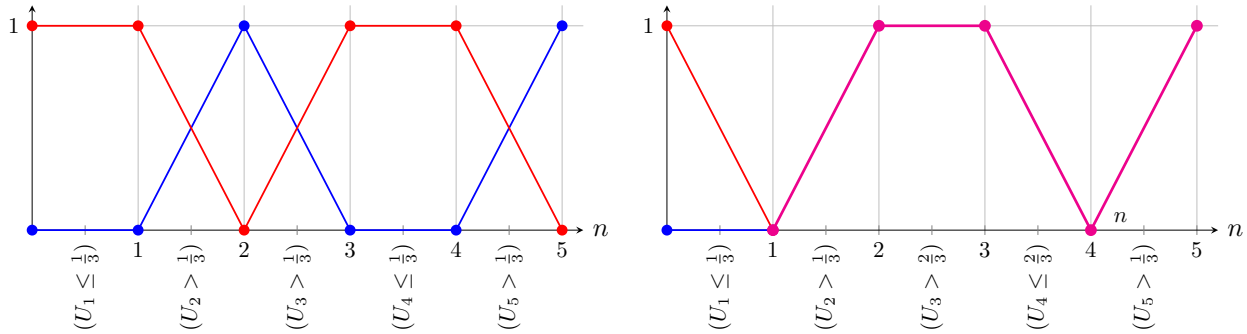


Figure 3: Markov chain of Example 1.3

Using this mapping, the two chains will never coalesce (see Figure 4a for example).



(a) Coalescence does not happen.

(b) Coalescence happens fast.

Figure 4: The choice of random mapping representation can change the coalescence time.

However, if we pick another random mapping representation

$$\Phi(0,u) = \begin{cases} 0 & \text{if } u \leq \frac{1}{3}, \\ 1 & \text{if } u > \frac{1}{3}, \end{cases} \quad \Phi(1,u) = \begin{cases} 0 & \text{if } u \leq \frac{2}{3}, \\ 1 & \text{if } u > \frac{2}{3}, \end{cases}$$

with the same realization of the U_n's, coalescence will take place in a few steps (see Figure 4b).

1.3 Coupling From The Past

Surprisingly perhaps, a slight modification of the idea of forward coupling leads to a criterion to check whether the chain is in the stationary distribution. The idea is called *coupling from the past* and is formalized as follows (the algorithm is known as the Propp-Wilson algorithm):

1. Generate once and for all $(U_{-n}, n \geq 1)$.
2. Set $m = +1$.
3. Start the experiment at all states $i \in S$ at time $-m$ and update $X_{n+1}^{(i,-m)} = \Phi\left(X_n^{(i,-m)}, U_{n+1}\right)$ for $n = -m, -m+1, \ldots, -1$ ($X_n^{(i,-m)}$ denotes the state of the chain at time n knowing that $X_{-m} = i$).
4. Check coalescence at time $n = 0$: If $X_0^{(i,-m)}$ is independent of i (i.e., $\forall i, j \in S,\ \ X_0^{(i,-m)} = X_0^{(j,-m)}$), $X_0^{(i,-m)}$ is the output and the algorithm terminates. If not, set $m \leftarrow m+1$ and return to step 3.

Let $T = \inf\{m \geq 1 : X_0^{(i,-m)} = X_0^{(j,-m)} \quad \forall i, j \in S\}$. In particular, it holds that $X_0^{(i,-T)} = X_0^{(j,-T)}$ for all $i, j \in S$. We can therefore rename this random variable as $X_0^{(-T)}$. We will see that under a very reasonable additional assumption, the distribution of $X_0^{(-T)}$ is *exactly* the stationary distribution π of the Markov chain.

Example 1.2 revisited for backward coupling. Let us compute in this case the probabilities $\mathbb{P}(X_0^{(-T)} = 0)$ and $\mathbb{P}(X_0^{(-T)} = 1)$. If the Propp-Wilson algorithm is correct, these two probabilities should be equal to those of the stationary distribution π, respectively 2/3 and 1/3.

As already observed, coalescence in this example can only take place in state 0. But now, we move backwards in time, which makes a difference. Suppose for example that until time $-(m-1)$, coalescence has not yet happened, i.e., $X_0^{(0,-(m-1))} \neq X_0^{(1,-(m-1))}$. Then the algorithm restarts from time $-m$, As we know that coalescence does not happen between time $-(m-1)$ and time 0 (remember that we use the same U_n's for each experiment), the only way it can happen is between time $-m$ and time $-(m-1)$, in which case necessarily $U_{-(m-1)} \leq 1/2$ and correspondingly, $X_{-(m-1)}^{(0,-m)} = X_{-(m-1)}^{(1,-m)} = 0$. Then the value of $X_0^{(0,-m)} = X_0^{(1,-m)}$ will depend on whether m is even or odd. If m is odd, this value will be 0; if m is even, this value will be 1.

There remains to determine the probability that coalescence takes place for an odd or even value of the starting time $-m$. Clearly, the probability that coalescence takes place for the first time when the algorithm starts at time $-m$ is 2^{-m} (as it must be that $U_{-(m-1)} \leq 1/2$ and $U_{-k} > 1/2$ for $0 \leq k \leq m-2$). So the probability that coalescence takes place with an odd starting time is $1/2 + 1/8 + 1/32 + \ldots = 2/3$, which is also the probability that $X^{(-T)} = 0$, and the probability that coalescence takes place with an even starting time is $1/4 + 1/16 + 1/64 + \ldots = 1/3$, which is also the probability that $X^{(-T)} = 1$. □

Let us now come back to the general case, and for $m, n \in \mathbb{Z}$ with $m \leq n$, define the event

$$A_{m,n} = \left\{ X_n^{(i,m)} = X_n^{(j,m)}, \forall i, j \in S \right\}$$

i.e., $A_{m,n}$ denotes the event that the chain coalesces between time instants m and n.

Theorem 1.4. Let $(X_n,\, n \geq 0)$ be an ergodic finite state Markov chain with limiting and stationary distribution π (and let $(X_n^{(i,m)},\, n \geq m)$ denote the same chain starting in state i at time m). If there exists $L > 0$ such that $\mathbb{P}(A_{0,L}) > 0$[1], then

(a) with probability 1, the Propp-Wilson algorithm outputs a value $X_0^{(-T)}$ in finite time;

(b) $X_0^{(-T)} \sim \pi$.

Proof. (a) In words, the algorithm terminates in finite time if and only if coalescence takes place. As there is by assumption a positive probability that coalescence takes place over a period of time of duration L, this ensures that coalescence takes place in finite time. In mathematical terms, we obtain that because X is a time-homogeneous Markov chain, the events $(A_{-kL,-(k-1)L},\, k \geq 1)$ are i.i.d., so

$$\mathbb{P}(\text{the algorithme terminates in finite time}) \geq \mathbb{P}\left(\bigcup_{k \geq 1} A_{-kL,-(k-1)L} \right)$$

$$= 1 - \mathbb{P}\left(\bigcap_{k \geq 1} \overline{A_{-kL,-(k-1)L}} \right) = 1 - \prod_{k \geq 1} \mathbb{P}\left(\overline{A_{-kL,-(k-1)L}} \right) = 1 - \lim_{n \to \infty} \left(1 - \mathbb{P}(A_{0,L})\right)^n = 1,$$

as $\mathbb{P}(A_{0,L}) > 0$ by assumption.

[1] If the chain is ergodic, then it is more than reasonable to assume that there exists a mapping Φ satisfying this assumption.

(b) Let μ be the distribution of $X_0^{(-T)} = X_0^{(i,-T)}$ for all $i \in S$. By time-homogeneity of the Markov chain,

$$X_0^{(i,-T)} \quad \text{and} \quad X_{-1}^{(i,-(T+1))} \quad \text{share the same distribution } \mu$$

Now, $X_0^{(i,-(T+1))} = \Phi(X_{-1}^{(i,-(T+1))}, U_0)$, so $X_0^{(i,-(T+1))} \sim \mu P$.

But by definition, T is the coalescence time, so it must be that $X_0^{(i,-(T+1))} = X_0^{(-T)}$ for all $i \in S$.

Therefore, $\mu P = \mu$, which proves that $\mu = \pi$ by uniqueness of the stationary distribution. □

1.3.1 Propp-Wilson Algorithm in Practice

In the Propp-Wilson algorithm depicted above, m is replaced by $m + 1$ at each iteration, so finding the coalescence time is a linear-time operation, which is too slow in practice. By replacing m with $2m$ instead, finding the coalescence time becomes a logarithmic-time operation.

Moreover, at first glance, the Propp-Wilson algorithm seems to be useless when the state space is huge (cf. some of the examples we have already seen in this course), as we need to keep track of $|S|$ copies of the chain in order to generate one sample distributed according to π. However, if we have

1. A partial ordering $\preceq$ on the state space S;
2. A *monotone* random mapping representation, i.e, a representation that preserves this ordering:

$$i \preceq j \Longrightarrow \Phi(i, u) \preceq \Phi(j, u) \quad \forall i, j \in S, \forall u \in [0, 1];$$

3. Two extremal states $\underline{i}$ and $\overline{i}$ such that $\underline{i} \preceq j \preceq \overline{i}$, $\forall j \in S$;

then we only need to keep track of the *two* chains $X^{(\underline{i},-m)}$ and $X^{(\overline{i},-m)}$. Indeed, under this assumption, all intermediary chains remain "sandwiched" between the two extreme chains as time goes by, so that at the time these two coalesce, all the others must have coalesced also.

Example 1.5. Consider the classical random walk on $S = \{0, 1, \ldots, N\}$, with transition probabilities

$$p_{00} = p_{01} = p_{N,N} = p_{N,N-1} = \frac{1}{2} \quad \text{and} \quad p_{i,i\pm1} = \frac{1}{2} \quad \text{for } i = 1, \ldots, N-1$$

and consider the following random mapping representation (with U_n i.i.d.$\sim \mathcal{U}[0, 1]$):

$$\begin{cases} \text{if } u \leq \frac{1}{2}: \quad \Phi(0, u) = 0, \quad \Phi(i, u) = i - 1 \quad \text{for } i = 1, \ldots, N \\ \\ \text{if } u > \frac{1}{2}: \quad \Phi(N, u) = N, \quad \Phi(i, u) = i + 1 \quad \text{for } i = 0, \ldots, N-1 \end{cases}$$

(Note that with such a mapping, the chain only possibly coalesces in states 0 or N). It is the case that this random mapping representation is monotone. So in this case, it sufficient to run the Propp-Wilson algorithm from initial states 0 and N only. Of course, there are more interesting examples one might be interested in! Next week, we will explore the Ising model.